AD-A100 442   OFFICE OF NAVAL RESEARCH ARLINGTON VA   F/G 12/2
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 28. NUMBER 2,(U)
JUN 81 S M SELIG

UNCLASSIFIED   NL

1 OF 2
AD A
100442

AD A100442

LEVEL II

1

6

NAVAL RESEARCH LOGISTICS QUARTERLY

DTIC
ELECTE
JUN 22 1981
S D
E

12 142

11

JUNE 1981
VOL. 28, NO. 2

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH

DTIC FILE COPY

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

81 6 18 030

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORIAL BOARD

Marvin Denicoff, *Office of Naval Research*, Chairman

Murray A. Geisler, *Logistics Management Institute*

W. H. Marlow, *The George Washington University*

Ex Officio Members

Thomas C. Varley, *Office of Naval Research*
Program Director

Seymour M. Selig, *Office of Naval Research*
Managing Editor

MANAGING EDITOR

Seymour M. Selig
Office of Naval Research
Arlington, Virginia 22217

ASSOCIATE EDITORS

Frank M. Bass, *Purdue University*
Jack Borsting, *Naval Postgraduate School*
Leon Cooper, *Southern Methodist University*
Eric Denardo, *Yale University*
Marco Fiorello, *Logistics Management Institute*
Saul I. Gass, *University of Maryland*
Neal D. Glassman, *Office of Naval Research*
Paul Gray, *Southern Methodist University*
Carl M. Harris, *Center for Management and Policy Research*
Arnoldo Hax, *Massachusetts Institute of Technology*
Alan J. Hoffman, *IBM Corporation*
Uday S. Karmarkar, *University of Chicago*
Paul R. Kleindorfer, *University of Pennsylvania*
Darwin Klingman, *University of Texas, Austin*
Kenneth O. Kortanek, *Carnegie-Mellon University*
Charles Kriebel, *Carnegie-Mellon University*
Jack Laderman, *Bronx, New York*
Gerald J. Lieberman, *Stanford University*
Clifford Marshall, *Polytechnic Institute of New York*
John A. Muckstadt, *Cornell University*
William P. Pierskalla, *University of Pennsylvania*
Thomas L. Saaty, *University of Pittsburgh*
Henry Solomon, *The George Washington University*
Wlodzimierz Szwarc, *University of Wisconsin, Milwaukee*
James G. Taylor, *Naval Postgraduate School*
Harvey M. Wagner, *The University of North Carolina*
John W. Wingate, *Naval Surface Weapons Center, White Oak*
Shelemyahu Zacks, *State University of New York at Binghamton*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

APPLICATIONS OF RENEWAL THEORY IN ANALYSIS OF THE FREE-REPLACEMENT WARRANTY*

n For
x
11.15
Codes
Dist | Special
A | 24

Wallace R. Blischke

University of Southern California
Los Angeles, California

Ernest M. Scheuer

California State University, Northridge
Northridge, California

ABSTRACT

Under a free-replacement warranty of duration W, the customer is provided, for an initial cost of C, as many replacement items as needed to provide service for a period W. Payments of C are not made at fixed intervals of length W, but in random cycles of length $Y = W + \gamma(W)$, where $\gamma(W)$ is the (random) remaining life-time of the item in service W time units after the beginning of a cycle. The expected number of payments over the life cycle, L, of the item is given by $M_Y(L)$, the renewal function for the random variable Y. We investigate this renewal function analytically and numerically and compare .he latter with known asymptotic results. The distribution of Y, and hence the renewal function, depends on the underlying failure distribution of the items. Several choices for this distribution, including the exponential, uniform, gamma and Weibull, are considered.

1. INTRODUCTION

Since a real or potential cost is involved, any item sold with a warranty must necessarily be priced higher than if it were sold without a warranty. How much more the seller should charge and how much more the buyer should be willing to pay depends upon the structure of the warranty and the life distribution of the item. An analysis of pro rata and free-replacement warranties from both buyer's and seller's points of view is given by Blischke and Scheuer [6] and [7].

In this paper we shall consider only the free-replacement warranty and shall be concerned mainly with the seller's (or supplier's, manufacturer's, and so forth) point of view. Of primary importance from this point of view is the long-run profitability of the item.

An important consideration in analyzing long-run profits for items sold under a free-replacement warranty is the expected income over the life cycle of the item. This, of course, is a function of the expected number of replacement items sold over the life cycle. This expected

*This research was supported by the Office of Naval Research under Contract No. N00014-75-C-0733, Task NR042-323 Code 434.

number, found from the renewal function for the associated random variable, is the subject of this investigation.

In the analysis it is assumed that the buyer purchases an identical replacement when the item in service at the end of the warranty period fails and that the purchase and initiation of operation of a replacement are instantaneous. It is also assumed that replacements are manufactured at the same cost and marketed at the same price. These are standard simplifying assumptions. Though obviously unrealistic, in practice they do not negate the results of the analysis because the important considerations are the cost/price relativities.

Another simplifying assumption made in the analysis is that the life cycle of the item is a constant. ("Life cycle" is also called "economic life" or "assumed life.") For planning purposes and for tax purposes, this is, indeed, customarily taken to be a fixed quantity. In reality, of course, equipment is purchased at different times and life cycles vary. Accordingly, the life cycle of the item could quite properly be considered to be a random variable. This, however, further complicates an already complex problem. Finally, it is not at all clear what might be reasonably realistic distributional assumptions. (We know of no studies that would suggest a particular distributional form.) Secondly, this would greatly complicate the renewal function.

It is suggested that in using the results of this paper, or any similar results, a parametric study be done, allowing L, W, C, g, etc. (defined below) to vary over some appropriate sets of values.

In the ensuing, we shall discuss in more detail the nature of the free-replacement warranty and its associated costs/profits, the role of renewal theory in analyzing warranty policies, and the specific renewal function encountered in the context just described.

The form of a renewal function depends ultimately on the underlying life distribution of the items in question. Typically in dealing with renewal functions, closed form expressions are available only for a few special cases, although limiting results are quite generally available. We shall find this to be true of the "special" renewal function under consideration here as well. Analytical results will be given for the exponential distribution and, to illustrate a point, the uniform distribution. Some results of a numerical investigation of the special renewal function for gamma and Weibull distributed lifetimes will also be discussed. These depend on a new analytical result and on newly calculated tables (details below).

2. THE SPECIAL RENEWAL FUNCTION AND ITS ROLE IN THE ANALYSIS OF WARRANTY POLICIES

The Analysis of Warranty Policies

In the analysis of warranty policies given by Blischke and Scheuer [6] and [7] the basic considerations were the comparison of cost to the consumer, and of profit to the supplier, of warranted versus unwarranted items. In the present paper, we shall limit attention to the point of view of the supplier. From his point of view, the cost comparison leads to the establishment of a differential pricing structure which will equate expected long-run profit in the two situations. Profit, of course, is a function of cost and income. In our previous work (Blischke and Scheuer [6]) we derived the expected profit *per warranty cycle*. Here we are concerned with the long-run profit over the life cycle of the item. This can be approximated for relatively long life cycles by pursuing an analysis along the lines of our 1975 paper [6]. (See especially Sections 2.1.1 and 2.2.) Our present objective is to obtain an exact expression for this quantity. A result of this type would provide a basis for evaluation of the adequacy of the approximation.

The Free-Replacement Warranty

The specific warranty policy under consideration here is the free-replacement policy. Under a warranty of this type the supplier provides replacements for failed items free of charge until a specified period of service, W, is attained. His income during this period is the price, C, charged for the initial item. His expected cost is the sum of the cost of supplying the initial item and the expected cost of all replacements required to provide the total warranted service time, W. In the sequel we shall express this expected cost following Blischke and Scheuer [6], as $g[1 + M_X(W)]$, where g is the cost per unit, X is the random lifetime of an individual item and $M_X(W)$ is the associated renewal function evaluated at W. (In this expression the quantity $1 + M_X(W)$ is the expected total number of items supplied; that is, the initial item plus the expected number of replacements.)

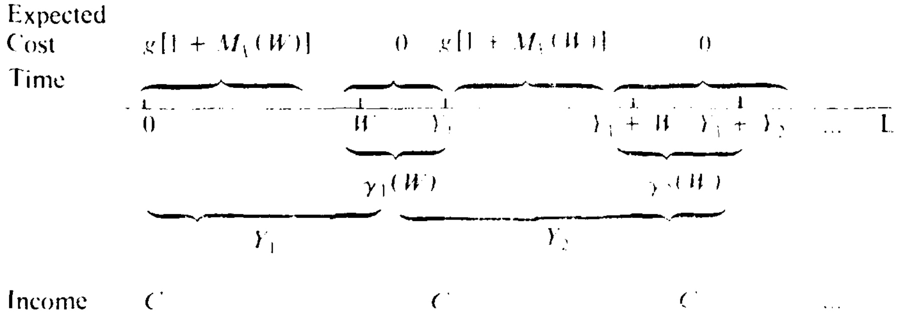
The Excess Random Variable

For the long-run analysis of the free-replacement warranty policy, it is important to note that no cost is incurred and no income obtained after W until the item in service at time W fails. The symbol Y is used to denote the random time at which this event takes place. This can also be expressed as $Y = W + \gamma(W)$, where $\gamma(W)$, the "excess random variable," is the (random) residual lifetime of the item in service at time W. This random variable is key to the analysis which follows. It is also called the "excess life" or "residual life" (Ross [13]), "remaining life" (Barlow and Proschan [2]), and "forward recurrence time" or "residual life-time" (Cox [9]), and has some unusual properties (see, for example, Feller [10]).

The Role of Renewal Functions

In the foregoing we have seen that the renewal function, $M_X(\cdot)$, of the basic lifetime random variable, X, plays an important role in determining expected profit on a per-cycle basis. In particular, expected profit per cycle is $P = C - g[1 + M_X(W)]$.

We turn now to the analysis of long-run expected profit. In this case we look at repetitions of the warranty cycle. The first such cycle extends from 0 to $Y_1 = W + \gamma_1(W)$, say; the second from Y_1 to Y_2; and so forth. Schematically, we have



The total expected profit is thus seen to be P times the number of expected repetitions of this process over the life cycle, L. This quantity is precisely the renewal function of the random variable Y, evaluated at L. We call this the special renewal function and denote it $M_Y(L)$. We can give a closed-form expression for $M_Y(L)$ for X having the exponential distribution and for L an integer multiple of W, Equation 18. Also, we can find explicitly the density and the moments of $\gamma(W)$ for X having the uniform distribution; however, the corresponding expression for $M_Y(\cdot)$ is not readily attainable, nor is a closed-form expression for $M_Y(\cdot)$, in general. However, asymptotic expressions for $M_Y(L)$ are available and some calculations, summarized in the portion of Section 4 showing results, indicate that a suitably chosen one of them can give quite satisfactory approximations to $M_Y(L)$ over a range of L values.

In our previous work we approximated $M_Y(L)$ by $L/E(Y)$. Our present numerical investigations indicate that this does not always provide an adequate approximation. By using a new renewal-theoretic result and with the aid of newly calculated tables we are able to obtain an improved, and altogether quite satisfactory, approximation (see Section 4).

3. ANALYTICAL INVESTIGATION OF $M_Y(\cdot)$

General Renewal-Theoretic Results

We begin with the basic renewal process involving a single warranty cycle. X, Y, $\gamma(\cdot)$, W and L are as defined previously. Let $X_1, X_2, \ldots$ be the lifetimes of the individual items within a warranty cycle. We assume that $X_1, X_2, \ldots$ are nonnegative random variables which are independent and identically distributed with cumulative distribution function $F_X(\cdot)$. We write $S_n = \sum_{i=1}^{n} X_i (n = 1, 2, \ldots), S_0 = 0$, $\mu = E(X)$, and $\sigma^2 = \text{var}(X)$. For any c.d.f., $F(\cdot)$, we define $\bar{F} = 1 - F$, $F^{(n)}(\cdot) =$ n-fold convolution of $F(\cdot)$ with itself, with [for $F(0-) = 0$]

$$F^{(0)}(t) = \begin{cases} 1 & t \geqslant 0 \\ 0 & t < 0. \end{cases}$$

In addition, we denote $N(t) =$ number of replacements required in the interval $(0, t]$, $M(t) = E(N(t))$, and $m(t) = M'(t)$.

A well-known, general renewal-theoretic result is that

$$P(N(t) = n) = F^{(n)}(t) - F^{(n+1)}(t). \tag{1}$$

This provides an immediate expression for $M(t)$ in terms of the convolutions $F^{(n)}(t)$, namely $M(t) = \sum_{n=1}^{\infty} F^{(n)}(t)$. We turn next to the problem of determining $F_Y(\cdot)$ and $F_Y^{(n)}(\cdot)$.

Many asymptotic results regarding renewal functions are available. Of primary interest here is the Elementary Renewal Theorem (Ross [13]), which was used in our previous work to approximate $M_Y(L)$. By this theorem, $M_Y(t)/t \rightarrow 1/E(Y)$ as $t \rightarrow \infty$. A further result, which we will exploit in the sequel, is (Cox [9]),

$$M_Y(t) \rightarrow \frac{t}{E(Y)} + \frac{\text{var}(Y)}{2E^2(Y)} - \frac{1}{2} \tag{2}$$

It has been known for some time (e.g., Smith [14]), that

$$E(Y) = \mu[1 + M_X(W)] \tag{3}$$

Recently Coleman [8] has found an expression for the moments of $\gamma(W)$, from which the moments of Y can be determined. In particular,

$$\text{var}(Y) = E(X^2)[1 + M_X(W)] - \mu^2[1 + M_X(W)]^2 + 2\mu\left[WM_X(W) - \int_0^W M_X(u)\,du\right]. \tag{4}$$

Coleman's result, along with newly calculated tables of $M_X(W)$ and $\int_0^W M_X(u)\,du$, permit the implementation of Equation (2). These tables will be described in Section 4.

Distribution of Y

Distribution of the Excess Random Variable

Since $Y = W + \gamma(W)$, the distribution of Y is simply a translation of the distribution of the excess random variable. Thus, the fundamental result required is the distribution of $\gamma(W)$. There are several ways of expressing this result. All, of course, relate back to the basic distribution of X since we can also write $\gamma(W)$ as $\gamma(W) = S_{N_X(W)+1} - W$.

The survival function for $\gamma(W)$ is given by Barlow and Proschan [2] as

$$\bar{F}_{\gamma(W)}(t) \equiv P\{\gamma(W) \geqslant t\} = \bar{F}_X(W + t) - \int_0^W \bar{F}_X(t + W - u)\,m_X(u)\,du. \tag{5}$$

An equivalent expression for the corresponding density is given by Cox [9] as

$$f_{\gamma(W)}(t) = f_X(W + t) + \int_0^W m_X(W - u)\,f_X(u + t)\,du. \tag{6}$$

Mixture Representation

It is of interest to note that in addition to these classical representations, the distribution of the excess random variable can also be expressed as a mixture of distributions (cf. Blischke [4] and [5]), namely

$$F_{\gamma(W)}(t) = \sum_{n=0}^{\infty} P\{\gamma(W) \leqslant t \mid N_X(W) = n\} P\{N_X(W) = n\}. \tag{7}$$

Here the distribution of N (given in Equation (1)) is the mixing distribution and the conditional distributions of γ given N are the components of the mixture. Since the event $\{N_X(W) = n\}$ is equivalent to the event $\{S_n < W, S_{n+1} \geqslant W\}$, the conditional distributions become

$$P\{\gamma(W) \leqslant t \mid N_X(W) = n\} = P\{S_{n+1} \leqslant W + t \mid S_n < W, S_{n+1} \geqslant W\}, \tag{8}$$

which can be expressed as an integral over the appropriate region of the bivariate distribution of S_n, S_{n+1}.

One property often encountered in dealing with mixed distributions is that they may be multimodal. This is indeed the case for the distribution of the excess random variable, a fact that became quite apparent in some of our computer simulations. Another property of mixtures of the type we are dealing with here is that the moments of the mixed distribution can be expressed as weighted averages of the moments of the components. We have not pursued this point but it would be of interest in some applications. (For example, one might be interested in the conditional expected residual lifetime of the item in service at the end of the warranty period, given that it is the nth replacement.)

An expression equivalent to Equation (7) is

$$\bar{F}_{\gamma(W)}(t) = \sum_{n=0}^{\infty} P\{\gamma(W) \geqslant t \cap N_X(W) = n\}. \tag{9}$$

In view of the remark preceding Equation (8) and using the definition of $\gamma(W)$, the joint probabilities in Equation (9) can, for $n \geqslant 1$, be written

$$\begin{aligned} P\{\gamma(W) \geqslant t \cap N_X(W) = n\} &= P\{S_{n+1} \geqslant t + W \cap S_n < W \cap S_{n+1} \geqslant W\} \\ &= P\{S_{n+1} \geqslant t + W \cap S_n < W\} \\ &= \int P\{t + W - u \leqslant S_n < W \mid X_{n+1} = u\} f_X(u)du && (10) \\ &= \int_t^{t+W} [F_X^{(n)}(W) - F_X^{(n)}(t + W - u)] f_X(u)du && (11) \\ &\quad + \int_{t+W}^{\infty} F_X^{(n)}(W) f_X(u)du. \end{aligned}$$

The limits of integration in Equation (11) come about as follows. In Equation (10) we require $t + W - u \leqslant W$, so $u \geqslant t$. Also if $t + W - u < 0$, i.e. $u > t + W$, then

$$(12) \qquad \begin{aligned} P\{t + W - u \leqslant S_n < W \mid X_{n+1} = u\} &= P\{0 \leqslant S_n < W \mid X_{n+1} = u\} \\ &= F_X^{(n)}(W), \end{aligned}$$

since we are dealing with nonnegative random variables. Also

$$(13) \qquad \begin{aligned} P\{\gamma(W) \geqslant t \cap N_X(W) = 0\} &= P\{X_1 > t + W \cap X_1 > W\} \\ &= P\{X_1 \geqslant t + W\} \\ &= \bar{F}_X(t + W). \end{aligned}$$

Using Equations (11) and (13) in Equation (9), we obtain

$$(14) \qquad \begin{aligned} \bar{F}_{\gamma(W)}(t) &= \bar{F}_X(t + W) + \sum_{n=1}^{\infty} \left[F_X^{(n)}(W) \int_t^{\infty} f_X(u)du \right. \\ &\qquad \left. - \int_t^{t+W} F_X^{(n)}(t + W - u) f_X(u)du\right] \\ &= \bar{F}_X(t + W) + M_X(W)\bar{F}_X(t) - \int_t^{t+W} M_X(t + W - u) f_X(u)du. \end{aligned}$$

Integrating by parts in Equation (14) and then making a change of variable in the resulting integral yields

$$(15) \qquad P\{\gamma(W) \geqslant t\} = \bar{F}_X(t + W) - \int_0^W \bar{F}_X(t + W - u) m_X(u)du,$$

which is Barlow and Proschan's formula cited at Equation (5) above.

The density for $\gamma(W)$ is, from Equation (14),

$$(16) \qquad \begin{aligned} f_{\gamma(W)}(t) &= -\frac{d}{dt} P\{\gamma(W) \geqslant t\} \\ &= f_X(t + W) + \int_t^{t+W} m_X(t + W - u) f_X(u)du \end{aligned}$$

which, by a change of variable of integration, is seen to be the same as Cox's formula cited at Equation (6) above.

To complete the analysis one has to pursue the derivation of the renewal function for Y. One approach is to translate the distribution of $\gamma(W)$ to obtain the distribution of Y, determine the n-fold convolution of this distribution with itself, and hence, by Equation (1), the distribution of N, and then determine $M = E(N)$ directly. Exact analytical expressions can be found by this approach only for a few special cases. In other cases the renewal function must be approximated, either by computer simulation or by using asymptotic results. The latter approach makes use of the Elementary Renewal Theorem or, better, of Equation (2).

Another approach to the determination of the renewal function of Y is via numerical integration. In principle, knowledge of $f_X(\cdot)$ permits calculation of $F_X(\cdot)$, of the $F_X^{(n)}(\cdot)$, and $M_X(\cdot)$. $F_{Y(W)}(\cdot)$ can be obtained from (2) [numerical differentiation of $M_X(\cdot)$ to get $m_X(\cdot)$ is needed here] and then the result $F_Y(t) = F_{Y(W)}(t - W)$ can be used. Then the successive convolutions, $F_Y^{(n)}(\cdot)$, can be calculated, from which, finally, $M_Y(\cdot)$ can be achieved. We have not attempted to implement this approach and know nothing about achievable accuracy or computing time requirements.

Examples

The Exponential Distribution

For the exponential distribution,

$$(17) \qquad f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0, \end{cases}$$

explicit expressions for all of the above are easily obtained. We use (6) to obtain the density of the excess random variable. The "renewal density" is $m(t) = 1/E(X) = \lambda$. Thus,

$$(18) \qquad \begin{aligned} f_{Y(W)}(t) &= \lambda e^{-\lambda(W+t)} + \int_0^W \lambda^2 e^{-\lambda(u+t)}du \\ &= \lambda e^{-\lambda t}, \qquad\qquad t > 0 \end{aligned}$$

which is, of course, a well-known result. The density of Y is simply a translated exponential. The n-fold convolution of this is a translated gamma distribution, with c.d.f.

$$(19) \qquad F_Y^{(n)}(y) = \begin{cases} 0 & y < nW \\ 1 - \sum\limits_{i=0}^{n-1} \dfrac{[\lambda(y - nW)]^i}{i!} e^{-\lambda(y-nW)} & y \geqslant nW \end{cases}$$

In writing the renewal function, it will be convenient to express L as an integer multiple of W, say $L = lW$. We then obtain, from Equations (1) and (19),

$$(20) \qquad P(N_Y(lW) = n) = e^{-\lambda(l-n-1)W} \sum_{i=0}^{n} \frac{\lambda^i(l-n-1)^i W^i}{i!} - e^{-\lambda(l-n)W} \sum_{i=0}^{n-1} \frac{\lambda^i(l-n)^i W^i}{i!}$$

$$n = 0, 1, \ldots, l-1$$

Finally, the special renewal function is found to be

$$(21) \qquad \begin{aligned} M_Y(lW) = E[N_Y(lW)] &= \sum_{j=1}^{l-1} j\{F_Y^{(j)}(lW) - F_Y^{(j+1)}(lW)\} \\ &= F_Y^{(1)}(lW) + F_Y^{(2)}(lW) + \ldots + F_Y^{(l-1)}(lW) \\ &\qquad - (l-1)F_Y^{(l)}(lW) \\ &= (l-1) - \sum_{j=1}^{l-1} e^{-j\lambda W} \sum_{i=0}^{j-1} \frac{(j\lambda W)^i}{i!} \end{aligned}$$

The Uniform Distribution

Although the uniform distribution is admittedly of limited interest as a life distribution, it is a convenient and nontrivial example to illustrate the mixture formulation. The density is

$$f_X(x) = \begin{cases} \dfrac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

It seems sensible to assume that $\theta > W$ since otherwise replacements are required with probability one. However, our analysis could easily be extended to cover the case $W > \theta$ with the formulas presented below.

The c.d.f. of the sum of n independent uniform $(0,\theta)$ random variables is

$$F_X^{(n)}(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > n\theta \\ \dfrac{1}{n!\theta^n}\left[x^n - \binom{n}{1}(x-\theta)^n + \binom{n}{2}(x-2\theta)^n - \dots + (-1)^k\binom{n}{k}(x-k\theta)^n\right] & \end{cases} \tag{23}$$

$$\text{for } k = 0, 1, \dots, n-1 \text{ and } k\theta \leqslant x \leqslant (k+1)\theta.$$

Recalling that $W < \theta$, we find directly that

$$\begin{aligned} P\{N_X(W) = n\} &= F_X^{(n)}(W) - F_X^{(n+1)}(W) \\ &= \frac{W^n}{n!\theta^{n+1}}\left[\theta - \frac{W}{n+1}\right]. \end{aligned} \tag{24}$$

Also, from Equation (23) and the fact that $M_X(x)$ can also be written as

$$M_X(x) = \sum_{n=1}^{\infty} F_X^{(n)}(x),$$

we find

$$1 + M_X(x) = \sum_{j=0}^{k} \frac{(-1)^j}{j!}\left(\frac{x - j\theta}{\theta}\right)^j \exp\left(\frac{x - j\theta}{\theta}\right), \quad k\theta \leqslant x \leqslant (k+1)\theta, \tag{25}$$

$$k = 0, 1, 2, \dots$$

The density of $\gamma(W)$ can be shown to be

$$f_{\gamma(W)}(t) = \begin{cases} \dfrac{1}{\theta}\, e^{\frac{W}{\theta}} & 0 < t < \theta - W \\ \dfrac{1}{\theta}\left[e^{\frac{W}{\theta}} - e^{\frac{t+W-\theta}{\theta}}\right] & \theta - W < t < \theta \\ 0 & \text{elsewhere.} \end{cases} \tag{26}$$

It follows from this that the distribution of $Y = W + \gamma(W)$ is

$$(27) \qquad f_Y(y) = \begin{cases} \dfrac{1}{\theta} e^{\frac{W}{\theta}} & W < y < \theta \\ \dfrac{1}{\theta}\left[e^{\frac{W}{\theta}} - e^{\frac{y-\theta}{\theta}}\right] & \theta < y < \theta + W \\ 0 & \text{elsewhere,} \end{cases}$$

with mean

$$(28) \qquad E(Y) = \frac{\theta}{2} e^{\frac{W}{\theta}}$$

and variance

$$(29) \qquad \sigma_Y^2 = e^{\frac{W}{\theta}}\left[\theta W - \frac{2\theta^2}{3}\right] + \theta^2 - \frac{\theta^2}{4} e^{\frac{2W}{\theta}} .$$

The above results can readily be used to express f_γ as a mixture. The mixing distribution is simply the distribution of N_W, given in Equation (24). The components of the mixture are conditional distributions, say $f_\gamma(\cdot|N_W(W) = n)$, of $\gamma(W)$ given $N_W(W) = n$. These are found to be

$$(30) \qquad f_\gamma(t|N_W(W) = n) = \begin{cases} \dfrac{1}{\theta - \dfrac{W}{n+1}} & 0 < t < \theta - W \\ \dfrac{W^n - (W + t - \theta)^n}{\theta W^n - \dfrac{W^{n+1}}{n+1}} & \theta - W < t < \theta. \end{cases}$$

In applications the conditional means of the excess random variable given N_W would also be of interest. Here we find

$$(31) \qquad E(\gamma(W)|N_W(W) = n) = \frac{\theta}{2} - \frac{\dfrac{W}{n+1}\left(\theta - \dfrac{2W}{n+2}\right)}{2\left(\theta - \dfrac{W}{n+1}\right)} .$$

The convolutions of $f_Y(\cdot)$ are rather tedious and we have not pursued this to get a closed expression for $M_Y(\cdot)$. One could, of course, use the Elementary Renewal Theorem with (28), or better, (2) with (28) and (29) to approximate $M_Y(\cdot)$. Finally, one might use an approach based on the result (Barlow & Proschan [2])

$$(32) \qquad M_Y^*(s) = \frac{F_Y^*(s)}{1 - F_Y^*(s)} ,$$

in which * denotes Laplace-Stieltjes transform, inverting to obtain $M_Y(\cdot)$.

The Gamma and Weibull Distributions

The gamma and Weibull distributions, with respective densities

$$(33)\qquad f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}}\, x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and

$$(34)\qquad f_X(x) = \begin{cases} \dfrac{\alpha}{\beta^{\alpha}}\, x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

are two of the more widely applied life distributions. Unfortunately, general, closed-form expressions for the basic renewal functions, $M_X(\cdot)$, to say nothing of the special renewal functions, $M_Y(\cdot)$, exist for neither. There is, however, a closed-form expression for the basic renewal function for the gamma distribution if the shape parameter, α, is integer-valued. (See, for example, Barlow and Proschan [1].) The renewal density for the gamma distribution with rational shape parameter can be obtained as well. (See Barlow and Proschan [2].) Series expressions for the renewal function for the Weibull distribution have been given by Smith and Leadbetter [15] and by Lomnicki [12]. Finally, the basic renewal function and other quantities have been evaluated for certain gamma and Weibull distributions by Soland [16], for the Weibull by White [17], for the lognormal, gamma, and Weibull by Huang [11] and for the gamma, inverse Gaussian, lognormal, truncated normal, and Weibull by Baxter, Scheuer, Blischke and McConalogue [3]. We will use various of these tabulations to aid us in approximating $M_Y(L)$ in Section 4.

4. NUMERICAL INVESTIGATION

Structure of the Numerical Studies

Because of the complexity encountered in the analytical investigation of the distribution of the excess random variable and the evaluation of the special renewal function, simulation programs were written to provide an opportunity to investigate the properties of both of these numerically. The basic life distributions that can be used in the simulations with these programs are the exponential, gamma, Weibull, uniform and normal. (The uniform for comparison with analytical results, the normal because of its apparent applicability in analyzing a set of data used as an example by Blischke and Scheuer [6], and the other three because they are the most important life distributions in the majority of applications.)

Here we shall concern ourselves only with the gamma and Weibull distributions. Some preliminary results concerning the special renewal function for these will be discussed below. The purpose of the special renewal program was to provide a means of investigating the approximation to $M_Y(L)/L$ using the asymptotic expression (2) and Equations (3) and (4).

The specific results which will be reported are for the following parameter combinations

α	β	
	Weibull	Gamma
2	1.12838	.500
3	1.11985	.333
4	1.10327	.250
5	1.08912	.200

These parameter combinations were initially chosen so that the tables of Soland [16] could be used to provide numerical values for the approximation. (Soland's tables are arranged to always have $\mu = 1$.) Subsequently, the new tables of Baxter, Scheuer, Blischke and McConalogue* became available and these were used in the calculations summarized in Tables 1 and 2, below. All combinations of $W = 0.5$, 1.0 and 1.5 with $L = 5$, 10, and 15 were used. (This gave warranty periods less than, equal to, and greater than the mean life and life cycles ranging from 3+ to 30 times the warranty period.) In each simulation 500 repetitions of the special renewal process were performed.

TABLE 1 — *Values of* $\hat{M}_Y(L)/L$, $1/E(Y)$, *and* $A(L)$ *for the Gamma Distribution*

	Parameters				$\hat{M}_Y(L)/L$			$A(L)$		
W	α	β	$1/E(Y)$	L:	5	10	15	5	10	15
0.5	2	1/2	.779		.707	.736	.753	.706	.743	.755
	3	1/3	.836		.750	.794	.805	.757	.797	.810
	4	1/4	.874		.792	.830	.840	.791	.833	.847
	5	1/5	.902		.819	.858	.868	.817	.859	.873
1.0	2	1/2	.570		.491	.526	.539	.484	.527	.541
	3	1/3	.601		.522	.561	.573	.511	.556	.571
	4	1/4	.618		.533	.578	.588	.527	.572	.587
	5	1/5	.628		.539	.587	.602	.536	.582	.598
1.5	2	1.2	.444		.350	.394	.413	.353	.399	.414
	3	1/3	.462		.369	.410	.428	.368	.415	.430
	4	1/4	.470		.377	.424	.440	.376	.423	.439
	5	1/5	.476		.380	.428	.443	.380	.428	.444

TABLE 2 — *Values of* $\hat{M}_Y(L)/L$, $1/E(Y)$, *and* $A(L)$ *for the Weibull Distribution*

	Parameters				$\hat{M}_Y(L)/L$			$A(L)$		
W	α	β	$1/E(Y)$	L:	5	10	15	5	10	15
0.5	2	1.13	.845		.757	.799	.814	.761	.803	.817
	3	1.12	.921		.832	.875	.888	.831	.876	.891
	4	1.10	.959		.869	.912	.927	.874	.916	.931
	5	1.09	.980		.882	.932	.947	.884	.932	.948
1.0	2	1.13	.616		.532	.574	.593	.525	.571	.586
	3	1.12	.653		.568	.612	.625	.560	.606	.622
	4	1.10	.666		.575	.622	.641	.572	.619	.635
	5	1.09	.676		.581	.633	.646	.587	.632	.646
1.5	2	1.13	.469		.372	.421	.436	.373	.421	.437
	3	1.12	.480		.388	.431	.446	.383	.432	.448
	4	1.10	.481		.396	.431	.449	.383	.432	.448
	5	1.09	.486		.397	.432	.453	.388	.437	.453

*These tables give $M_X(t)$, $\operatorname{var}[N_X(t)]$, and $\int_0^t M_X(u)du$ for X having gamma, inverse Gaussian, lognormal, truncated normal, and Weibull distributions, they encompass a broad range of parameter values and of values of t. We note that Soland's 1968 tables do not explicitly give $\int_0^t M_X(u)du$, but do include the variance of the associated equilibrium renewal process, $V(t) = \frac{t}{\mu} - \left(\frac{t}{\mu}\right)^2 + \frac{2}{\mu}\int_0^t M_X(u)du$, from which values of the integral can easily be obtained.

Results

In each of the simulations the average number of renewals, say $\hat{M}_Y(L)$ was calculated (along with certain additional relevant summary statistics). The basic results for the gamma distribution are given in Table 1 and for the Weibull distribution in Table 2. In each case the values tabulated are $\hat{M}_Y(L)/L$. For comparison purposes, values of $1/E(Y)$ are included, as well as values of the asymptotic approximation of $\frac{1}{E(Y)} + \frac{1}{L}\left[\frac{\text{var}(Y)}{2E^2(Y)} - \frac{1}{2}\right] = A(L)$.

In the simulations we also calculated the sample variances of the number of renewals for the random variable Y. From these results one can estimate the standard error of $\hat{M}_Y(L)/L$. The results ranged from less than .002 to .009, with all standard errors except those for combinations of the smallest values of W and L less than .005. Given that the accuracy of the computer simulations themselves is adequate, one can therefore conclude that we have the second digit determined to within one unit or so, except for a few cases.

Discussion

It is important to note that the approximation based on the Elementary Renewal Theorem is somewhat inaccurate: $1/E(Y)$ always overestimates $\hat{M}_Y(L)/L$, with the difference, of course, decreasing as L increases. (Thus $L/E(Y)$ would consistently overestimate $\hat{M}_Y(L)$ which would lead to an overestimate of the expected income over the life cycle of the item.)

The asymptotic approximation $A(L)$ gives quite good agreement with $\hat{M}_Y(L)/L$. The relative discrepancy between these two quantities occasionally runs up to 2%, but is mostly well below 1%. Accordingly, it is apparent that $LA(L)$ will generally provide a satisfactory approximation to $M_Y(L)$ — certainly so in the absence of an exact mathematical expression for $M_Y(L)$ or tables of that quantity.

ACKNOWLEDGMENTS

We thank Professor R. M. Soland for bringing the thesis of his student, C. N. Huang, to our attention and for providing us with a copy. We thank Dr. R. Coleman for making a copy of his paper [8] available to us prior to its publication. We also thank the Editor and a referee for helpful comments. The support of the Office of Naval Research is gratefully acknowledged.

REFERENCES

[1] Barlow, R.E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, Inc., New York, N.Y., 1965).

[2] Barlow, R.E. and F. Proschan, *Statistical Theory of Reliability and Life Testing, Probability Models* (Holt, Rinehart and Winston, Inc., New York, N.Y., 1975).

[3] Baxter, L.A., E.M. Scheuer, W.R. Blischke and D.J. McConalogue, "Renewal Tables: Tables of Functions Arising in Renewal Theory," technical report, School of Business Administration, University of Southern California, to appear (1981).

[4] Blischke, W.R., "Mixtures of Discrete Distributions," in *Classical and Contagious Discrete Distributions*, G.P. Patil, Editor (Statistical Publishing Society, Calcutta, India, 1965, distributed by Pergamon Press).

[5] Blischke, W.R., "Distributions, Statistical. IV. Mixtures of Distributions," in *International Encyclopedia of the Social Sciences*, Vol. IV, D.L. Sills, Editor (Crowell Collier and MacMillan, Inc., New York, N.Y., 1968).

[6] Blischke, W.R. and E.M. Scheuer, "Calculation of the Cost of Warranty Policies as a Function of Estimated Life Distributions," Naval Research Logistics Quarterly, *22*, 4, 681-696 (1975).

[7] Blischke, W.R. and E.M. Scheuer, "Application of Nonparametric Methods in the Statistical and Economic Analysis of Warranties," in *The Theory and Applications of Reliability, with Emphasis on Bayesian and Nonparametric Methods*, Vol. II, C.P. Tsokos and I.N. Shimi, Editors (Academic Press, Inc., New York, N.Y., 1977).

[8] Coleman, R., "The Moments of Forward Recurrence Time," submitted for publication. (Copies of this paper may be obtained by writing Dr. R. Coleman, Math. Dept., Imperial College, London SW7 2BZ, England.)

[9] Cox, D.R., *Renewal Theory* (Methuen and Co., Ltd., London, 1962).

[10] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. II (John Wiley and Sons, Inc., New York, N.Y., 1966).

[11] Huang, C.N., "The Numerical Computation of Renewal Functions," Masters Thesis, University of Texas at Austin (1972).

[12] Lomnicki, Z.A., "A Note on the Weibull Renewal Analysis," Biometrika, *53*, 375-381 (1966).

[13] Ross, S.M., *Applied Probability Models with Optimization Applications* (Holden-Day, Inc., San Francisco, CA, 1970).

[14] Smith, W.L., "Renewal Theory and Its Ramifications," Journal of the Royal Statistical Society, *20B*, 243-302 (1958).

[15] Smith, W.L. and M.R. Leadbetter, "On the Renewal Function for the Weibull Distribution," Technometrics, *5*, 393-396 (1963).

[16] Soland, R.M., "Renewal Functions for Gamma and Weibull Distributions with Increasing Hazard Rate," Technical Paper RAC-TP-329, Research Analysis Corp., McLean, VA (1968).

[17] White, J.S., "Weibull Renewal Analysis," *Proceedings of the Third Annual Aerospace Reliability and Maintainability Conference*, 639-657 (Society of Automotive Engineers, New York, N.Y., 1964).

COMPARING ALTERNATING RENEWAL PROCESSES

Dalen T. Chiang

College of Business Administration
Cleveland State University
Cleveland, Ohio

Shun-Chen Niu

School of Management
University of Texas at Dallas
Richardson, Texas

ABSTRACT

Sufficient conditions are given for stochastic comparison of two alternating renewal processes based on the concept of uniformization. The result is used to compare component and system performance processes in maintained reliability systems.

1. INTRODUCTION AND SUMMARY

Comparison of stochastic processes has been a rapidly growing area of research. In this paper, we will study alternating renewal processes (ARP) $X = \{X(t),\ t \geqslant 0\}$ where the state space $S = \{0,\ 1\}$ and the holding times of the process in state 1 and 0 are independent random variables having distribution functions F and G. Throughout this paper, we assume F and G are absolutely continuous with failure rate functions $r(t)$ and $q(t)$, respectively. We shall denote such a process by $(X,\ r(t),\ q(t))$. Similar notations will be used throughout.

Let $X = \{X(t),\ t \in T\}$ and $Y = \{Y(t),\ t \in T\}$ be two stochastic processes. We say X is *stochastically larger* than Y, denoted by $X \overset{st}{\geqslant} Y$, iff $E\,f(X) \geqslant E\,f(Y)$ for all nondecreasing functionals f for which the expectations exist. If X and Y have the same distribution, then we write $X \overset{st}{=} Y$. In a recent paper, Sonderman [8] presented a set of sufficient conditions such that stochastic comparison between two semi-Markov processes can be made. By specializing his conditions to the case of alternating renewal processes, Sonderman (Theorem 5.1 of [8]) obtained the following result.

THEOREM 1 (Sonderman): Let $(X^i,\ r_i(t),\ q_i(t))$, $i = 1,\ 2$, be two alternating renewal processes. Assume that time 0 is a renewal point for both processes and

(a) $X^1(0) \overset{st}{\leqslant} X^2(0)$,

(b) $r_1(u) \geqslant r_2(v)$,

(c) $q_1(u) \leqslant q_2(v)$,

for all $u, v \geqslant 0$, then there exist two ARP's $\hat{X}^1$ and $\hat{X}^2$ defined on the same probability space Ω such that $\hat{X}^i \overset{st}{=} X^i$, $i = 1, 2$, and $\hat{X}^1 \leqslant \hat{X}^2$ everywhere in Ω.

The purpose of this note is to show that conditions (b) and (c) in Theorem 1 can be weakened to

(b') $r_1(u) \geqslant r_2(v)$ whenever $u \leqslant v$,

(c') $q_1(v) \leqslant q_2(u)$ whenever $u \leqslant v$.

The proof of this result and two immediate corollaries will be presented in Section 2. Section 3 contains some remarks on the main results.

2. PATHWISE COMPARISON OF ALTERNATING RENEWAL PROCESSES

We shall start by describing a construction due to Sonderman [8] which reproduces an alternating renewal process $(X, r(t), q(t))$ based on a Poisson process. In order to do that, the following technical assumption on $r(t)$ and $q(t)$ is needed.

ASSUMPTION: The alternating renewal process $(X, r(t), q(t))$ is assumed to be *uniformizable*, i.e., there exists a real number $\lambda < \infty$ such that $\sup_{t \geq 0} \{r(t), q(t)\} \leqslant \lambda$. λ is called the *uniformization rate*.

As discussed in Sonderman [8, pp. 113-115], this condition can be relaxed to the case where failure rates are uniformly bounded over finite intervals. Let λ be the uniformization rate of X, the construction can be separated into two steps. First, a Poisson process with rate λ generates a sequence of potential transition epochs $\{t_i, i \geqslant 0\}$, where $t_0 \equiv 0$. Then a discrete time stochastic process is constructed on $\{t_i, i \geqslant 0\}$, determining whether each potential transition epoch is a genuine transition and, if so, the new state of the process. Specifically, let $\{(S_n, J_n), n \geqslant 0\}$ be a sequence of ordered pairs of integer-valued random variables, where S_n has the value 1 or 0 representing the state of the process immediately after t_n. The variable $J_n = m (m \leqslant n)$ if the last genuine transition is at t_m. We assume a genuine transition occurs at $t = 0$, i.e., $J_0 = 0$. The initial state $S_0 = X(0)$ could either be given or have an initial probability distribution. The transition probabilities of $\{(S_n, J_n), n \geqslant 0\}$ are defined as

(1)
$$
\begin{aligned}
&P(S_n = 0, J_n = n \mid S_{n-1} = 1, J_{n-1} = m, t_i, i \geqslant 0) = r(t_n - t_m)/\lambda \\
&P(S_n = 1, J_n = n \mid S_{n-1} = 0, J_{n-1} = m, t_i, i \geqslant 0) = q(t_n - t_m)/\lambda \\
&P(S_n = S_{n-1}, J_n = J_{n-1} \mid S_{n-1}, J_{n-1}, t_i, i \geqslant 0) \\
&\qquad = 1 - P(J_n = n \mid S_{n-1}, J_{n-1}, t_i, i \geqslant 0) \text{ for } 0 \leqslant m < n.
\end{aligned}
$$

Finally, define a new process $\hat{X} = \{\hat{X}(t), t \geqslant 0\}$ by

(2)
$$\hat{X}(t) = S_n \quad \text{if} \quad t_n \leqslant t < t_{n+1}.$$

Then it follows from Theorem 2.1 of Sonderman [8] that $\hat{X} \overset{st}{=} X$.

We will need the following lemma from Arjas and Lehtonen ([1], Lemma 3). See also Theorem 3.1 of [8].

LEMMA 1: Let $X = \{X_n, n \geqslant 0\}$, $Y = \{Y_n, n \geqslant 0\}$, and $Z = \{Z_n, n \geqslant 0\}$ be three discrete time stochastic processes. Suppose that

(a) $(X_0 | Z_n = z_n, n \geqslant 0) \overset{st}{\leqslant} (Y_0 | Z_n = z_n, n \geqslant 0)$

and (b) $(X_j | X_0 = x_0, \ldots, X_{j-1} = x_{j-1}, Z_n = z_n, n \geqslant 0) \overset{st}{\leqslant}$

$(Y_j | Y_0 = y_0, \ldots, Y_{j-1} = y_{j-1}, Z_n = z_n, n \geqslant 0)$

whenever $x_i \leqslant y_i$, $0 \leqslant i \leqslant j-1$, for all $j \geqslant 1$. Then there exist two stochastic processes $\hat{X} = \{\hat{X}_n, n \geqslant 0\}$ and $\hat{Y} = \{\hat{Y}_n, n \geqslant 0\}$ defined on the same probability space such that $\hat{X} \overset{st}{=} X$, $\hat{Y} \overset{st}{=} Y$, and $\hat{X} \leqslant \hat{Y}$ everywhere, hence, $X \overset{st}{\leqslant} Y$.

We are now ready to state and prove the main theorem of this paper.

THEOREM 2: Let $(X^i, r_i(t), q_i(t))$, $i = 1, 2$, be two uniformizable alternating renewal processes. Assume that time 0 is a renewal point for both processes and

(a) $X^1(0) \overset{st}{\leqslant} X^2(0)$,

(b′) $r_1(u) \geqslant r_2(v)$ whenever $u \leqslant v$,

(c′) $q_1(v) \leqslant q_2(u)$ whenever $u \leqslant v$

then there exist two new processes $\hat{X}^1$ and $\hat{X}^2$ defined on the same probability space Ω such that $\hat{X}^1 \overset{st}{=} X^1$, $\hat{X}^2 \overset{st}{=} X^2$ and $\hat{X}^1 \leqslant \hat{X}^2$ everywhere in Ω, hence $X^1 \overset{st}{\leqslant} X^2$.

PROOF: The proof is a modification of the one used by Sonderman [8] to prove his Theorem 3.2. Since both processes are Poisson-uniformizable, let $\lambda \geqslant 2 \sup_{t \geqslant 0} \{r_1(t), q_2(t)\}$. The basic idea of the proof is to generate potential transition epochs for both processes by the same Poisson process. Let $\{t_n, n \geqslant 0\}$ be a sequence of events generated by a Poisson process with rate λ. In view of Lemma 1, we need only to show that the two discrete time stochastic processes $\{S_n^1, n \geqslant 0\}$ and $\{S_n^2, n \geqslant 0\}$ constructed according to (1) and (2) from X^1 and X^2, respectively, satisfy the following stochastic order relationships:

$$(S_j^1 | S_0^1 = s_0^1, \ldots, S_{j-1}^1 = s_{j-1}^1, t_n, n \geqslant 0) \overset{st}{\leqslant}$$

$$(S_j^2 | S_0^2 = s_0^2, \ldots, S_{j-1}^2 = s_{j-1}^2, t_n, n \geqslant 0)$$

whenever $s_i^1 \leqslant s_i^2$, $0 \leqslant i \leqslant j-1$ for all $j \geqslant 1$, or equivalently,

$$P(S_j^1 = 1 | S_0^1 = s_0^1, \ldots, S_{j-1}^1 = s_{j-1}^1, t_n, n \geqslant 0) \leqslant \tag{3}$$

$$P(S_j^2 = 1 | S_0^2 = s_0^2, \ldots, S_{j-1}^2 = s_{j-1}^2, t_n, n \geqslant 0)$$

whenever $s_i^1 \leqslant s_i^2$, $0 \leqslant i \leqslant j-1$ for all $j \geqslant 1$.

Suppose $(s_0^1, \ldots, s_{j-1}^1) \leqslant (s_0^2, \ldots, s_{j-1}^2)$, and let $J_{j-1}^1 = k^1$ and $J_{j-1}^2 = k^2$, where $0 \leqslant k^1 \leqslant j-1$ and $0 \leqslant k^2 \leqslant j-1$.

CASE 1: Suppose $s^1_{l+1} = 1$, hence, $s^2_{l+1} = 1$.

In this case, $k^1 \geq k^2$ and $t_l - t_{k^1} \leqslant t_l - t_{k^2}$. Then by (1) and condition (b′),

left hand side of (3) $= 1 - r_1(t_l - t_{k^1})/\lambda \leqslant 1 - r_2(t_l - t_{k^2})/\lambda$ = right-hand side of (3).

CASE 2: Suppose $s^1_{l+1} = 0$ and $s^2_{l+1} = 1$.

l.h.s. of (3) $= q_1(t_l - t_{k^1})/\lambda \leqslant 1/2 \leqslant 1 - r_2(t_l - t_{k^2})/\lambda$ = r.h.s. of (3)

CASE 3: Suppose $s^1_{l+1} = s^2_{l+1} = 0$

In this case, $k^1 \leqslant k^2$ and $t_l - t_{k^1} \geq t_l - t_{k^2}$. Then from (1) and condition (c′), we have

l.h.s. of (3) $= q_1(t_l - t_{k^1})/\lambda \leqslant q_2(t_l - t_{k^2})/\lambda$ = r.h.s. of (3).

The conclusion of the theorem now follows from Lemma 1 since

$$S^1_0 = X^1(0) \overset{st}{\leqslant} X^2(0) = S^2_0.$$

Q.E.D.

The following corollaries are immediate.

COROLLARY 1: Conditions (a), (b′), and (c′) in Theorem 2 can be replaced by

(i) $X^1(0) \overset{st}{\leqslant} X^2(0)$,

(ii) $r_1(t)$ or $r_2(t)$ is nonincreasing in t,

(iii) $q_1(t)$ or $q_2(t)$ is nonincreasing in t,

(iv) $r_1(t) \geq r_2(t)$ and $q_1(t) \leqslant q_2(t)$ for all $t \geq 0$.

PROOF: Suppose $u \leqslant v$. If $r_1(t)$ is nonincreasing, then $r_1(u) \geq r_1(v) \geq r_2(v)$. If $r_2(t)$ is nonincreasing, then $r_1(u) \geq r_2(u) \geq r_2(v)$. Hence, in either case, condition (b′) of Theorem 2 is satisfied. Condition (c′) can be checked in similar fashion.

Q.E.D.

COROLLARY 2: Let $(X, r(t), q(t))$ be a uniformizable alternating renewal process. Then there exist two alternating renewal processes $(Y, r_Y(t), q_Y(t))$ and $(Z, r_Z(t), q_Z(t))$, where

$$r_Z(t) = \sup_{t \leq s < \infty} r(s), \qquad q_Z(t) = \inf_{0 \leq s \leq t} q(s),$$

$$r_Y(t) = \inf_{0 \leq s \leq t} r(s), \qquad q_Y(t) = \sup_{t \leq s < \infty} q(s),$$

such that X is bounded stochastically from below by Z and from above by Y.

PROOF: Clearly the functions $r_Z(t)$, $q_Z(t)$, $r_Y(t)$, and $q_Y(t)$ are non-increasing in t. Therefore, the conclusion is a direct consequence of Corollary 1.

Q.E.D.

3. COMMENTS AND ADDITIONS

(1) In Theorem 2, the assumption that time 0 is a renewal point for both processes can be relaxed. It is sufficient to assume that at time 0, if both processes are in state 1, then X^2 has been in state 1 longer than X^1, and if both processes are in state 0, then X^1 has been in state 0 longer than X^2.

(2) In a loose sense, the processes Z and Y in Corollary 2 may be viewed as the greatest lower bound and least upper bound, respectively, for process X within the class of alternating renewal processes whose holding times in both states are DFR (decreasing failure rate).

(3) An alternating renewal process may be used to model the performance of a repairable component in a maintained reliability system (see [3] or Chapter 6 of [2]). The successive operating (or repair) times of a repairable component are assumed to be independent and identically distributed random variables. All components operate independently of one another. Let $X(t)$ be the state of a component at time t, where

$$X(t) = \begin{cases} 1 & \text{if the component is up at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

then $X = \{X(t),\ t \geqslant 0\}$ is an alternating renewal process. Therefore, Theorem 2 may be used to compare the performance of two maintained reliability systems consisting of n repairable components. Specifically, let ϕ be a coherent structure function (see [2]) and $X_i^j = \{X_i^j(t),\ t \geqslant 0\}$ be the performance process of the ith component in jth systems, where $i = 1, 2, \ldots, n$, $j = 1, 2$. Define $\underline{X}^j(t) = (X_1^j(t), \ldots, X_n^j(t))$, $j = 1, 2$. By forming the product of probability spaces for individual components, the following result follows directly from Theorem 2.

PROPOSITION 1: Suppose that

(i) $X_i^1(0) \leqslant X_i^2(0)$ for all $i = 1, \ldots, n$.

(ii) All component performance processes are uniformizable and the failure rates satisfy the conditions of Theorem 2.

Then there exist two stochastic processes $\hat{\phi}^1$ and $\hat{\phi}^2$ defined on the same probability space Ω such that $\hat{\phi}^1 \overset{st}{=} \{\phi(\underline{X}^1(t)),\ t \geqslant 0\}$, $\hat{\phi}^2 \overset{st}{=} \{\phi(\underline{X}^2(t)),\ t \geqslant 0\}$, and $\hat{\phi}^1 \leqslant \hat{\phi}^2$ everywhere in Ω. Hence, $\{\phi(\underline{X}^1(t)),\ t \geqslant 0\} \overset{st}{\leqslant} \{\phi(\underline{X}^2(t)),\ t \geqslant 0\}$.

(4) It is interesting to point out that an example of Miller [5, example (ii), p. 308] shows that increasing the failure rate of downtime distribution of a component does not necessarily increase (stochastically) the time to first system failure or system availability. Our result (see Corollary 1) shows that for systems whose repairable components have DFR uptime and downtime distributions, decreasing the failure rates of uptime distributions and increasing the failure rates of downtime distributions do improve the system performance.

(5) Theorem 2 may be used to establish bounds for performance measures of maintained reliability systems. For example, one can bound the performance process of a repairable component by that of a component whose uptime and downtime distributions are exponential (This is a special case of Corollary 1 here or Theorem 5.1 of [8]). Maintained systems with exponential uptime and downtime distributions has been discussed in Brown [4], Ross [6 and 7]. However, the bounds obtained in this fashion are usually quite loose. Finally, we present the following example to illustrate the ideas involved:

EXAMPLE: Consider a two-component parallel system. Let $F(G)$ be the uptime (downtime) distribution of component 1 and $\lambda(\mu)$ be the constant failure (repair) rate for component 2. Assume the system starts operation with both components new. Suppose we are interested in the expected time until first system failure, $E(T_0)$. By conditioning on the state of the second component when component 1 fails for the first time, it is not difficult to see that

$$E(T_0) = \int_0^\infty t\,dF(t) + \left[\int_0^\infty P_{11}(t)\,dF(t)\right]\cdot\left[E(\min[D,U]) + \left(\int_0^\infty e^{-\lambda y}\,dG(y)\right)E(T_0)\right]$$

where $D(U)$ is a random variable having distribution G (exponential distribution with parameter λ) and $P_{11}(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}$. After some simplification, we have

$$E(T_0) = \frac{\int_0^\infty t\,dF(t) + \left[\int_0^\infty P_{11}(t)\,dF(t)\right]\cdot\left[\int_0^\infty e^{-\lambda y}(1-G(y))\,dy\right]}{1 - \left[\int_0^\infty P_{11}(t)\,dF(t)\right]\cdot\left[\int_0^\infty e^{-\lambda y}\,dG(y)\right]} \equiv h(F,G;\lambda,\mu).$$

Therefore, we can [illegible] bounds for $E(T_0)$ for a two-component parallel system whose first component has the same performance process as above and the second component performance process is uniformizable with failure rate function $\lambda(t)$ and repair rate function $\mu(t)$, $t \geq 0$. Specifically, let $\bar{\lambda} = \sup_{t \geq 0}\{\lambda(t)\}$, $\underline{\lambda} = \inf_{t \geq 0}\{\lambda(t)\}$, $\bar{\mu} = \sup_{t \geq 0}\{\mu(t)\}$, and $\underline{\mu} = \inf_{t \geq 0}\{\mu(t)\}$, then

$$h(F,G;\bar{\lambda},\underline{\mu}) \leqslant E(T_0) \leqslant h(F,G;\underline{\lambda},\bar{\mu}).$$

REFERENCES

[1] Arjas, E. and T. Lehtonen, "Approximating Many Server Queues by Means of Single Server Queues," Mathematics of Operations Research, *3*, 205-223 (1978).

[2] Barlow, R.E. and F. Proschan, *Statistical Theory of Reliability and Life Testing: Probability Models* (Holt, Rinehart, and Winston, New York, N.Y., 1975).

[3] Barlow, R.E. and F. Proschan, "Theory of Maintained Systems: Distribution of Time to First System Failure," Mathematics of Operations Research, *1*, 32-42 (1976).

[4] Brown, M., "The First Passage Time Distributions for a Parallel Exponential System with Repair," in R.E. Barlow, J.B. Fussell and N. Singpurwalla, Editors, *Reliability and Fault Tree Analysis* (SIAM, Philadelphia, 1975).

[5] Miller, D.R., "A Continuity Theorem and Some Counterexamples for the Theory of Maintained Systems," Stochastic Processes and Their Applications, *5*, 307-314 (1977).

[6] Ross, S.M., "On Time to First Failure in Multicomponent Exponential Reliability Systems," Journal of Stochastic Processes and Their Applications, *4*, 167-173 (1976).

[7] Ross, S.M. and J. Schechtman, "On the First Time a Separately Maintained Parallel System Has Been Down for a Fixed Time," Naval Research Logistic Quarterly, *26*, 285-290 (1979).

[8] Sonderman, D., "Comparing Semi-Markov Processes," Mathematics of Operations Research, *5*, 110-119 (1980).

SHOCK MODELS WITH PHASE TYPE SURVIVAL AND SHOCK RESISTANCE

Marcel F. Neuts*

University of Delaware
Newark, Delaware

Manish C. Bhattacharjee**

Indian Institute of Management
Calcutta, India

ABSTRACT

New closure theorems for shock models in reliability theory are presented If the number of shocks to failure and the times between the arrivals of shocks have probability distributions of phase type, then so has the time to failure PH-distributions are highly versatile and may be used to model many qualitative features of practical interest. They are also well-suited for algorithmic implementation The computational aspects of our results are discussed in some detail

1. INTRODUCTION

Shock models which relate the life distribution $H(\cdot)$ of a device, subject to failure by shocks occurring randomly in time, have received considerable attention in recent years. If $\bar{P}_k$ is the probability that the device survives $k \geq 0$, shocks and $N(t)$ is the random number of shocks in $(0,t]$, the survival probability, $\bar{H}(\cdot) = 1 - H(\cdot)$, of such a device is given by

$$(1) \qquad \bar{H}(t) = E\bar{P}_{N(t)} = \sum_{k=0}^{\infty} \bar{P}_k \; P\{N(t) = k\}.$$

The most general shock models are those that correspond to (1), such that $\{N(t);\ t \geq 0\}$ is a general counting process and $1 \geq \bar{P}_0 \geq \bar{P}_1 \geq \bar{P}_2 \geq \ldots$. Interest in and published results for shock models center around proving that, subject to suitable assumptions on the point process $N(t)$ of shocks, various reliability characteristics of the shock resistance probabilities $\bar{P}_k$ are inherited by the survival probability $\bar{H}(\cdot)$ in continuous time.

The first systematic treatment of such shock models was given by Esary, Marshall and Proschan [5], when $N(t)$ is a homogeneous Poisson process. A-Hameed and Proschan considered the cases when $N(t)$ is a nonhomogeneous Poisson process [1] and a nonstationary

*This research was supported by the National Science Foundation under Grant No ENG-7908351 and by the Air Force Office of Scientific Research under Grant No AFOSR-77-3236.

**This research was partially supported by research project 441/CMDS-APR-1 at the Indian Institute of Management Calcutta

pure birth process [2]. Block and Savits [4] treated the case when the interarrival time between shocks is NBUE (NWUE) or NBU (NWU) and Thall [8] derived interesting, but comparatively weaker, results when $N(t)$ is a clustered Poisson process.

In this paper, we obtain preservation theorems for the shock model (1) when $\bar{P}_k$ is of *phase-type* and so is the distribution of the interarrival time between shocks. $N(t)$ is then a phase type renewal process [7]. The relevance of phase type distributions (henceforth abbreviated as PH-distributions) to the algorithmic analysis of the time dependent behavior of stochastic models has been discussed by Neuts in a series of papers starting with [6]. A comprehensive treatment may be found in Chapter 2 of [8]. PH-distributions provide an alternative point of departure in modelling real life distributions without the classic *memoryless property and* with possible proper unimodality or multimodality. PH-distributions include the exponential, Erlang and hyperexponential distributions as very special cases. In addition, they have the desirable property of being closed under both finite convolutions and mixtures, a feature possessed by none of the well-known nonparametric classes of life distributions.

In Section 2, the basic properties of PH-distributions, needed in the sequel, are briefly reviewed. The main theoretical results are discussed in Section 3. Algorithmic considerations are presented in Section 4.

2. PH-DISTRIBUTIONS

A density $\{p_k\}$ on the nonnegative integers is *of phase type* if and only if there exists a finite Markov chain with transition probability matrix P of order $r+1$ of the form

$$P = \begin{vmatrix} S & \underline{S}^{o} \\ \underline{0} & 1 \end{vmatrix},$$

and initial probability vector $[\underline{\beta}, \beta_{r+1}]$, such that $\{p_k\}$ is the density of the time till absorption in the state $r+1$. The matrix $I - S$ is nonsingular and the stochastic matrix $S + (1 - \beta_{r+1})^{-1}$ $\underline{S}^{o} \cdot \underline{\beta}$ may be chosen to be irreducible.

The density $\{p_k\}$ is given by $p_0 = \beta_{r+1}$, and $p_k = \underline{\beta}\, S^{k-1}\, \underline{S}^{o}$, for $k \geqslant 1$. In this paper $\{p_k\}$ will be the density of the number of shocks to failure in a reliability shock model. We will assume throughout that $\beta_{r+1} = 0$. We also clearly have that

$$\bar{P}_k = \sum_{j=k+1}^{\infty} p_j = \underline{\beta} S^k \underline{e}, \qquad \text{for } k \geqslant 0.$$

The mean μ_1' of $\{p_k\}$ is given by $\underline{\beta}(I - S)^{-1}\underline{e}$.

A probability distribution $F(\cdot)$ on $[0, \infty)$ is *of phase type* if and only if there exists a finite Markov process with generator Q of the form

$$Q = \begin{vmatrix} T & \underline{T}^{o} \\ \underline{0} & 0 \end{vmatrix},$$

with initial probability vector $[\underline{\alpha}, \alpha_{m+1}]$, such that $F(\cdot)$ is the distribution of the time till absorption in the state $m+1$. The matrix T is nonsingular and the generator $T + (1 - \alpha_{m+1})^{-1}\, \underline{T}^{o} \cdot \underline{\alpha}$ may be chosen to be irreducible. The distribution $F(\cdot)$ is given by

$$(2) \qquad F(x) = 1 - \underline{\alpha} \exp(Tx)\underline{e}, \qquad \text{for } x \geqslant 0.$$

We shall denote $1 - F(x)$ by $\bar{F}(x)$. The mean λ_1' of $F(\cdot)$ is given by $\lambda_1' = -\underline{\alpha}\, T^{-1}\underline{e}$. The pairs $(\underline{\alpha}, T)$ and $(\underline{\beta}, S)$ are called *representations* of $F(\cdot)$ and $\{p_k\}$ respectively. Renewal processes in which the underlying distribution $F(\cdot)$ is of phase type were discussed in [7].

Many derivations related to PH-distributions involve the Kronecker product $L \otimes M$ of two matrices L and M. This is the matrix made up of the blocks $\{L_{ij}M\}$. Provided the matrix products are defined, we have that

$$(L \otimes M)\,(K \otimes H) = LK \otimes MH. \tag{3}$$

This property is repeately used in the sequel.

3. CLOSURE THEOREMS

We first consider the Esary-Marshall-Proschan (E.M.P.) shock model [3,5] in which $\{N(t)\}$ is a Poisson counting process of rate λ.

THEOREM 1

If the number of shocks to failure has a discrete PH-density $\{p_k, k \geqslant 0\}$ with representation $(\underline{\beta}, S)$, then the time to failure in the E.M.P. model has a continuous PH-distribution $H(\cdot)$ with representation $[\underline{\beta}, \lambda(S - I)]$.

PROOF

Since $\bar{P}_k = \underline{\beta}\, S^k \underline{e}$, for $k \geqslant 0$, we obtain

$$\bar{H}(t) = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!}\, \underline{\beta}\, S^k \underline{e} = \underline{\beta} \exp\,[\lambda(S - I)t]\, \underline{e}, \quad \text{for } t \geqslant 0.$$

This proves the stated result.

A number of interesting quantities may now be expressed in computationally convenient forms. The j-th noncentral moment of $H(\cdot)$ is given by

$$\mu_j' = j!\,\lambda^{-j}\underline{\beta}(I - S)^{-j}\underline{e}, \quad \text{for } j \geqslant 1. \tag{4}$$

The density $h(t) = H'(t)$, is given by

$$h(t) = \lambda\, \underline{\beta} \exp\,[\lambda(S - I)t]\, \underline{S}^{\circ}, \quad \text{for } t \geqslant 0, \tag{5}$$

and the failure rate $r(t) = h(t)\bar{H}^{-1}(t)$, equals

$$r(t) = \lambda \frac{\underline{\beta} \exp\,(\lambda\, t\, S)\underline{S}^{\circ}}{\underline{\beta} \exp\,(\lambda\, t\, S)\,\underline{e}}, \quad \text{for } t \geqslant 0. \tag{6}$$

Theorem 1 is a particular case of a more general result in which the arrivals of shocks occur according to a PH-renewal process [7]. This result is proved next.

Let the interarrival time distribution $F(\cdot)$ be of phase type with irreducible representation $(\underline{\alpha}, T)$ of order m. When $\alpha_{m+1} = 1 - \underline{\alpha}\,\underline{e}$, is positive, a geometrically distributed number of shocks occur simultaneously at each shock epoch. As in [7], we introduce the matrices $P(k,t)$, $k \geqslant 0$, $t \geqslant 0$, which satisfy the system of differential equations

(7) $$P'(0,t) = P(0,t)T,$$

$$P'(k,t) = P(k,t)T + \sum_{\nu=1}^{k} \alpha_{m+1}^{\nu-1} P(k-\nu,t)\underline{T}^{\circ}\underline{\alpha}, \qquad k \geqslant 1,$$

for $t \geqslant 0$, with initial conditions $P(k,0) = \delta_{0k}\, I$, for $k \geqslant 0$. The element $P_{ij}(k,t)$ is the conditional probability that the Markov process with generator $Q^* = T + (1-\alpha_{m+1})^{-1}\, \underline{T}^{\circ}\underline{\alpha}$, is in the state j at time t and that k shocks have occurred in $(0,t]$, given that it started in the state i at time 0.

The Markov process Q^* may be started according to any initial probability vector $\underline{\gamma}$. With $\underline{\gamma} = (1-\alpha_{m+1})^{-1}\underline{\alpha}$, the PH-renewal process is started immediately after a renewal epoch. With $\underline{\gamma} = -\lambda_1'^{-1}\,\underline{\alpha}T^{-1}$, where $\lambda_1' = -\underline{\alpha}T^{-1}\underline{e}$, is the mean time between shocks, we obtain the stationary version of the PH-renewal process.

THEOREM 2

If the shocks occur according to a PH-renewal process with underlying representation $(\underline{\alpha},T)$ and the process Q^* is started according to the probability vector $\underline{\gamma}$ and if the probability density $\{p_k\}$ is of phase type with representation $(\underline{\beta},S)$ of order r, then the distribution $H(\cdot)$ is of phase type with the representation

(8) $$\underline{\kappa} = \underline{\gamma} \otimes \underline{\beta},$$

$$K = T \otimes I + \underline{T}^{\circ}\underline{\alpha} \otimes (I - \alpha_{m+1}S)^{-1}S,$$

of order rm.

PROOF

By the law of total probability, we have

(9) $$\bar{H}(t) = \underline{\gamma} \sum_{k=0}^{\infty} P(k,t)\underline{e} \cdot \underline{\beta}S^k\underline{e}$$

$$= (\underline{\gamma} \otimes \underline{\beta}) \sum_{k=0}^{\infty} P(k,t) \otimes S^k\, (\underline{e} \otimes \underline{e})$$

$$= (\underline{\gamma} \otimes \underline{\beta})\; Z(t)\; (\underline{e} \otimes \underline{e}), \qquad \text{for } t \geqslant 0.$$

The matrix $Z(t) = \sum_{k=0}^{\infty} P(k,t) \otimes S^k$, satisfies

$$Z'(t) = \sum_{k=0}^{\infty} P'(k,t) \otimes S^k = \sum_{k=0}^{\infty} P(k,t)\; T \otimes S^k$$

$$+ \sum_{k=1}^{\infty} \sum_{\nu=1}^{k} \alpha_{m+1}^{\nu-1}\, P(k-\nu,t)\; \underline{T}^{\circ}\underline{\alpha} \otimes S^k$$

$$= Z(t)\;(T \otimes I) + \sum_{k=0}^{\infty} P(k,t)\; \underline{T}^{\circ}\alpha \otimes S^{k+1}\,(I - \alpha_{m+1}S)^{-1}$$

$$= Z(t)\;[T \otimes I + \underline{T}^{\circ}\underline{\alpha} \otimes (I - \alpha_{m+1}S)^{-1}S],$$

and clearly $Z(0) = I \otimes I$.

This implies that $Z(t) = \exp(Kt)$, for $t \geqslant 0$. Upon substitution into (9), the proof is complete.

Particular Cases

1. If the number of shocks to failure is geometrically distributed, i.e., $\bar{P}_k = \theta^k$, for $k \geqslant 0$, $0 < \theta < 1$, then

$$\bar{H}(t) = \underline{\gamma} \sum_{k=0}^{\infty} P(k,t)\theta^k \underline{e} = \underline{\gamma} \exp\{[T + (1 - \theta\alpha_{m+1})^{-1}\theta \underline{T}^0 \underline{\alpha}]t\} \underline{e}, \tag{10}$$

for $t \geqslant 0$.

2. In the *maximum shock model*, failure occurs if and only if a shock occurs whose magnitude exceeds a critical randomized threshold Y with distribution $G(\cdot)$. If the magnitudes of successive shocks are independent with common distribution $F(\cdot)$, then

$$\bar{P}_k = \int_0^{\infty} F^k(x)\, dG(x), \qquad \text{for } k \geqslant 0. \tag{11}$$

It follows from (10) that

$$\bar{H}(t) = \int_0^{\infty} \underline{\gamma} \exp\{[T + (1 - \alpha_{m+1}F(x))^{-1} F(x)\underline{T}^0 \underline{\alpha}]t\} \underline{e}\, dG(x), \tag{12}$$

for $t \geqslant 0$, so that $H(\cdot)$ is a mixture of PH-distributions. If $G(\cdot)$ is a discrete distribution with finite support, then $H(\cdot)$ itself is of phase type. Case 1 above corresponds to $G(\cdot)$ being degenerate at θ.

3. In *the cumulative damage model*, the damages are additive. With the same distributions $F(\cdot)$ and $G(\cdot)$ as in the preceding model, we obtain

$$\bar{P}_k = \int_0^{\infty} F^{(k)}(x)\, dG(x), \qquad \text{for } k \geqslant 0. \tag{13}$$

If the distribution $G(\cdot)$ is of phase type with representation $(\underline{\delta}, L)$ and $X_1, \ldots, X_k$ are i.i.d. with common distribution $F(\cdot)$, then

$$\bar{P}_k = \int_0^{\infty} \bar{G}(x)\, dF^{(k)}(x) = E\bar{G}(X_1 + \ldots + X_k)$$

$$= E\,\underline{\delta} \exp[L(X_1 + \ldots + X_k)]\underline{e} = \underline{\delta} A^k \underline{e},$$

where $A = \int_0^{\infty} \exp(Lx)\, dF(x)$. It is readily seen that A is a substochastic matrix of spectral radius less than one. The density $\{p_k\}$ is therefore of phase type. If the shocks occur according to a PH-renewal process, Theorem 2 may be applied to evaluate $\hat{H}(t)$. The matrix A is obtained by numerical integration for general distributions $F(\cdot)$. If $F(\cdot)$ itself is of phase type with representation $(\underline{\sigma}, R)$, then

$$\begin{aligned} A &= \int_0^{\infty} \exp(Lx)\underline{\sigma} \exp(Rx)\underline{R}^0\, dx \\ &= (I \otimes \underline{\sigma}) \int_0^{\infty} \exp(Lx) \otimes \exp(Rx)\, dx\, (I \otimes \underline{R}^0) \\ &= -(I \otimes \underline{\sigma})[L \otimes I + I \otimes R]^{-1}(I \otimes \underline{R}^0). \end{aligned} \tag{14}$$

The eigenvalues of L and R all lie in the open left half-plane. The same then holds true for the Kronecker sum $L \otimes I + I \otimes R$, so that the inverse exists.

The nonnegative rectangular matrix $V = -(L \times I + I \times R)^{-1}(I \times \underline{R}^0)$, may easily be computed by solving the system

$$(L \otimes I + I \otimes R)V = -I \otimes \underline{R}^0$$

by block Gauss-Seidel iteration

4. ALGORITHMIC ASPECTS

We shall discuss the computation of the function $\bar{H}(t)$, which is given by Theorem 2. It readily follows from (1) that the mean h_1' of $H(\cdot)$ is given by $\lambda_1'\mu_1'$, where λ_1' and μ_1' are the means of $\{p_k\}$ and $F(\cdot)$ respectively, whenever the PH-renewal process of arrivals is started at a renewal epoch. With general initial conditions, the mean h_1' is given by $\lambda_1'\mu_1' + \tilde{\lambda}_1 - \lambda_1'$, where $\tilde{\lambda}_1 = -\gamma T^{-1}\underline{e}$.

Knowledge of the mean h_1' of $H(\cdot)$ is useful in determining the interval over which we wish to evaluate $\bar{H}(t)$. We may, e.g., wish to choose the mean as a convenient unit of time. This is accomplished by replacing K by $h_1'K$. A different rescaling may be chosen if the elements of $h_1'K$ are very large or if a different time scale is desirable for the practical problem at hand.

We now assume that the matrix K has been appropriately rescaled. The function $\bar{H}(t)$ is computed by numerical integratrion of the system of linear differential equations

$$\underline{v}'(t) = \underline{v}(t)K, \qquad \text{for } t \geqslant 0, \tag{15}$$

$$\underline{v}(0) = \gamma \otimes \underline{\beta},$$

and setting $\bar{H}(t) = \underline{v}(t)\underline{e}$, for $t \geqslant 0$.

It is convenient to partition the vector $\underline{v}(t)$ as $[\underline{v}_1(t), \ldots, \underline{v}_m(t)]$, where the vectors $\underline{v}_i(t)$ are r-vectors. We also set $M = (I - \alpha_{m+1}S)^{-1}S$. The system (15) may then be rewritten as

$$\underline{v}_j'(t) = \sum_{i=1}^{m} \underline{v}_i(t)\, T_{ij} + \alpha_j \left[\sum_{r=1}^{m} \underline{v}_r(t)\, T_r^{\circ}\right] M, \tag{16}$$

for $1 \leqslant j \leqslant m$. This system may be conveniently solved by a classical integration procedure, such as Runge-Kutta. We see that the vector $\left[\sum_{r=1}^{m} \underline{v}_r(t)\, T_r^{\circ}\right] M$ does not depend on j and needs to be evaluated only once in each computation of the right-hand sides of (16).

In many PH-distributions of practical interest, such as, e.g., finite mixtures of Erlang distributions, the order m of T may be large, but T, $\underline{T}^{\circ}$ and $\underline{\alpha}$ have very few nonzero entries. It is then advantageous to write a special purpose subroutine to evaluate the right-hand side of (16). By so exploiting the sparsity of T, $\underline{T}^{\circ}$ and $\underline{\alpha}$, it is possible to reduce the computation time greatly. The mean h_1', or in general the scaling factor used in selecting the time unit, may also be utilized to choose the step size h in the numerical integration of the system (16). In similar problems, we have usually made two runs at least, one with 1/50 of the time unit and one with 1/100 of the time unit. If the results at corresponding time points are not sufficiently close, further runs with smaller steps are made. The computation times of such runs increase rapidly and efficient programming is desirable. Other methods with a variable step size and error control may also be implemented. These classical topics in the numerical integration of ordinary differential equations need not be belabored here. In all cases, the use of the particular structure of the matrix K is fully worthy of the additional programming effort.

BIBLIOGRAPHY

[1] A-Hameed, M.S. and F. Proschan, "Non-stationary Shock Models," Stochastic Processes and Their Applications, *1*, 383-404 (1973).

[2] A-Hameed, M.S. and F. Proschan, "Shock Models with Underlying Birth Process," Journal of Applied Probability, *12*, 18-28 (1975).

[3] Barlow, R.E. and F. Proschan, *Statistical Theory of Reliability and Life Testing: Probability Models* (Holt, Rinehart and Winston, New York, N.Y. 1975).

[4] Block, H.W. and T.H. Savits, "Shock Models with NBUE Survival," Journal of Applied Probability, *15*, 621-628.

[5] Esary, J.D., A.W. Marshall, and F. Proschan, "Shock Models and Wear Processes," Annals of Probability, *1*, 627-649 (1973).

[6] Neuts, M.F. "Probability Distributions of Phase Type," in *Liber Amicorum Professor Emeritus H. Florin*, 173-206, Department of Mathematics, University of Louvain, Belgium (1975).

[7] Neuts, M.F., "Renewal Processes of Phase Type," Naval Research Logistics Quarterly, 25, 445-454 (1978).

[8] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models—An Algorithm Approach* (The Johns Hopkins University Press, Baltimore, MD (1981).

[9] Thall, P.F. "Cluster Shock Models," Tech. Rept. No. 47, University of Texas at Dallas, Dallas, TX (1979).

AN EARLY-ACCEPT MODIFICATION TO THE TEST PLANS OF MILITARY STANDARD 781C

David A. Butler*

Oregon State University
Corvallis, Oregon

Gerald J. Lieberman**

Stanford University
Stanford, California

ABSTRACT

This paper is concerned with the statistical test plans contained in Military Standard 781C, "Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution" and the selection and use of these plans. Modifications to the fixed-length test plans of MIL-STD-781C are presented which allow early-accept decisions to be made without sacrificing statistical validity. The proposed plans differ from the probability ratio sequential tests in the Standard in that rejection is permitted only after a fixed number of failures have been observed.

1. INTRODUCTION AND SUMMARY

Military Standard 781C, "Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution" [2] covers the requirements for reliability qualification tests (pre-production) and reliability acceptance tests (production) for equipment that experiences a distribution of times-to-failure that is exponential. These requirements include: test conditions, procedures, and various fixed-length and sequential test plans with respective accept/reject criteria. This paper is concerned only with the statistical test plans and the selection and use of these plans. The Standard contains both fixed-length test plans (Plans IXC through XVIIC and XIXC through XXIC) and probability-ratio sequential tests (Plans IC through VIIIC and XVIIIC). Each fixed-length test plan is characterized by its discrimination ratio (d), its total test time (T), and its maximum allowable number of failures to accept (k). If a fixed-length test plan is selected, the total test duration is essentially set in advance. The only way in which one of these plans can terminate early is by rejection. For example, Test Plan XVIIC terminates with a reject decision at the third failure if this failure occurs before 4.3 units of total test time have transpired. An accept decision can only be made when 4.3 units of total test time have accrued. Even if the second failure occurs very early, an early reject decision cannot be made; nor can an early-accept decision be made if no failures have occurred,

*This research was supported in part by Contract N00014-79-C-0751 with the Office of Naval Research

**This research was supported in part by Contract N00014-75-C-0561 with the Office of Naval Research.

say, by time 4.0. In both of these situations, an early decision would lack statistical validity in failing to guarantee the operating characteristic of the selected plan. Moreover, an early reject decision by the consumer would probably violate contractual agreements with the producer. However, an early-accept decision by the consumer would not be subject to such an objection. Such a decision might seem very desirable to the consumer (government) if testing costs were substantial or if schedule deadlines were near. This paper presents modifications to the fixed-length test plans of MIL-STD-781C which allow early-accept decisions to be made without sacrificing statistical validity. The proposed plans differ from the probability ratio sequential tests in the Standard in that rejection is permitted only after a fixed number of failures have been observed.

2. THE EARLY-ACCEPT CRITERION

The early-accept criterion we will consider is as follows. Consider a test plan $\mathcal{P}_k$ with discrimination ratio d, total test time T_k, maximum allowable number of failures to accept $k\ (k \geq 1)$, and consumer's risk β. Consider alternative test plans $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{k-1}$ with the same discrimination ratio, maximum allowable number of failures to accept $j\ (0 \leq j < k)$, and total test times $T_j = \frac{1}{2}\cdot\chi^2_{(1-\beta;\,2j+2)}$, where $\chi^2_{(1-\beta;\,2j+2)}$ is the $(1-\beta)$ percentile of a chi-squared distribution with $2j+2$ degrees of freedom.* The producer's risks for test plans $\mathcal{P}_j\ (0 \leq j \leq k)$ are in decreasing order of j, the test times are in increasing order of j, and the consumer's risks are constant in j (each is β).

The early-accept criterion is as follows: accept at time T_j if at most j failures have occurred up to that time. The reject criterion remains as before: reject at the $(k+1)$st failure. The early-accept modification alters the original test plan $\mathcal{P}_k$ by allowing early-accept decisions *to be made at k time points prior to* the total test time T_k. As a result the producer's risk for test plan $\mathcal{P}_k$ is altered. Also, even though each test plan $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_k$ has consumer's risk β, and even though the alternative test plans $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{k-1}$ were only involved with accept decisions, the consumer's risk of the resulting test is not maintained at β, and indeed, may be significantly greater than β. It is true that if an early-accept decision is made at time T_j, then test plan $\mathcal{P}_j$, had it been selected prior to the start of testing, would have reached the same conclusion. But, by allowing the test results to effectively dictate which test plan is used, the probability calculations involved in determining the consumer's risk are modified by the conditional probabilities which must consequently be incorporated into them. The producer's and consumer's risks for the modified test plans are computed as follows. Let $P_A(\lambda)$ denote the probability of accepting when the true mean time between failures (MTBF) is $1/\lambda$.

$$P_A(\lambda) = \sum_{j=0}^{k} Pr\{\text{accept at time } T_j\}.$$

Let $A(j) = Pr\{\text{accept at time } T_j\}$.

THEOREM 1: Suppose the true MTBF is $1/\lambda$. Then

$$A(j) = \frac{(\lambda T_j)^j \exp(-\lambda T_j)}{j!} - \sum_{i=0}^{j-1} A(i)\cdot\frac{[\lambda(T_j - T_i)]^{j-i}\exp(-\lambda(T_j - T_i))}{(j-i)!}.$$

PROOF: If an accept decision is made at time T_j, then exactly j failures must have occurred up to that time (since if fewer than j failures had occurred, an accept decision would have been made earlier). Thus,

*This choice is somewhat arbitrary, but is motivated by the use of this rule to guarantee a given consumer's risk for a fixed-length test plan.

$$Pr\{\text{exactly } j \text{ failures in } [0, T_j]\} = Pr\left\{ \biguplus_{l=0}^{j} [\text{accept at time } T_l \text{ and } (j-l) \text{ failures in } (T_l, T_j]] \right\},$$

where $\uplus$ represents a union of disjoint events.

$$\frac{(\lambda T_j)^j \exp(-\lambda T_j)}{j!} = \sum_{l=0}^{j-1} A(l) \cdot \frac{[\lambda(T_j - T_l)]^{j-l} \exp(-\lambda(T_j - T_l))}{(j-l)!} + A(j).$$

The consumer's risk for the early-accept test plan is $P_t(1)$ and the producer's risk is $1 - P_t(1/d)$.

3. EARLY-ACCEPT TEST PLANS

It has been proposed that the early-accept criterion be used with the existing parameters of the fixed-length test plans of MIL-STD-781C. The effect of incorporating the early-accept criterion into these fixed-length test plans (without further modification) is shown in Table 1. In all plans except Plan XXIC the consumer's risk is increased and the producer's risk is decreased. (Test Plan XXIC is unchanged since it only accepts when there are no failures.) The changes are substantial; often the consumer's risk is more than doubled and the producer's risk halved. By altering the test time and the maximum nu mber of failures to accept, it is possible to correct for the effect of the early-accept modification and closely match the operating characteristics (at two points) of the standard fixed-length test plans. The corrections for each of the MIL-STD-781C fixed-length test plans are given in Table 2. Accept times for these early-accept test plans are listed in Table 3.

The corrections were computed by defining functions $f_\alpha(T,k)$ as the producer's risk for an early-accept test plan with parameters T and k, and $f_\beta(T,k)$ as the consumer's risk. As T increases f_α increases and f_β decreases, and as k increases f_α decreases and f_β increases. Because of the integer restriction on k, it is not always possible to design a test plan to achieve specified values of α, β exactly. However, an algorithm which will determine an approximate solution can be constructed. The algorithm from which Table 2 is derived first fixes k and uses a quasi-Newton method to determine a value of T which will achieve the desired α-value. The process is then repeated, varying k in accordance with a bisection search, to determine a k-value for which β is also close to the desired level. Some additional checks to reduce the calculations are also incorporated. It should be noted that the test plans of Table 2 are designed to have α and β levels close to the nominal values of the standard test plans, not the actual values. (See Tables II and C-I in [2]).

4. PERFORMANCE OF THE EARLY-ACCEPT TEST PLANS

Table 2 shows that the maximum test times for the early-accept test plans are substantially increased from the standard test times. However, the expected test times for the early-accept plans are much smaller than the maximum times, and compare quite favorably to the (fixed) test times for the standard plans.* Graphs of expected test duration versus true MTBF for the early-accept test plans appear in Figures 1-12. For comparison, the figures also graph the expected test duration versus true MTBF for the standard test plans. The early-accept plans

*The expected test times for Early-Accept Plans IXC and XC exceed those for the corresponding standard plans for a considerable range of the true MTBF. The reason for this is that these two early-accept plans have producer's and consumer's risks substantially closer to the nominal values than do the standard plans.

TABLE 1 — *Changes in Producer's and Consumer's Risks Resulting from Incorporating Early-Accept Criterion into MIL-STD-781C Test Plans*

Test Plan	Discrimination Ratio	Without Early-Accept Option*		With Early-Accept Option‡	
		Producer's Risk (%)	Consumer's Risk (%)	Producer's Risk (%)	Consumer's Risk (%)
IXC	1.5	12.0	9.9	4.9	38.1
XC	1.5	10.9	21.4	3.5	58.8
XIC	1.5	17.8	22.1	6.8	56.4
XIIC	2.0	9.6	10.6	4.7	31.8
XIIIC	2.0	9.8	20.9	4.4	48.4
XIVC	2.0	19.9	21.0	11.3	42.8
XVC	3.0	9.4	9.9	5.9	23.1
XVIC	3.0	10.9	21.3	6.8	38.4
XVIIC	3.0	17.5	19.7	12.5	32.6
(High Risk Plans)					
XIXC	1.5	28.8	31.3	14.0	59.5
XXC	2.0	28.8	28.5	19.4	44.6
XXIC	3.0	30.7	33.3	30.7	33.3

*Taken from Tables II and III of MIL-STD-781C and is for the test plan *without* early-accept modification

‡True risk when the early-accept criterion is incorporated

TABLE 2 — *Specifications of Standard and Early-Accept Test Plans*

Test Plan	Discrimination Ratio	MIL-STD-781C Test Plans**					
		Test Time*	No. of Failures to Reject	Test Time*	No. of Failures to Reject	Producer's Risk for Corrected Plan (%)‡	Consumer's Risk for Corrected Plan (%)
IXC	1.5	45.0	≥ 37	72.2	≥ 55	10.2	10.0
XC	1.5	29.9	≥ 26	51.7	≥ 40	10.1	19.8
XIC	1.5	21.1	≥ 18	32.6	≥ 24	20.1	20.4
XIIC	2.0	18.8	≥ 14	26.0	≥ 17	10.4	10.3
XIIIC	2.0	12.4	≥ 10	19.1	≥ 13	9.9	19.2
XIVC	2.0	7.8	≥ 6	12.6	≥ 8	20.0	18.3
XVC	3.0	9.3	≥ 6	12.8	≥ 7	10.0	8.4
XVIC	3.0	5.4	≥ 4	8.3	≥ 5	10.2	18.7
XVIIC	3.0	4.3	≥ 3	5.2	≥ 3	19.7	19.2
High Risk Plans							
XIXC	1.5	8.0	≥ 7	12.6	≥ 9	29.6	30.8
XXC	2.0	3.7	≥ 3	4.5	≥ 3	29.9	29.1
XXIC	3.0	1.1	≥ 1	1.1	≥ 1	30.7	33.3

*In multiples of θ_1

**From Tables II and III in MIL-STD-781C

‡Corrected for use with early-accept criterion to achieve true producer's and consumer's risks close to nominal levels as given in Table C-1 of MIL-STD-781C

TABLE 3 — *Accept Times of Early-Accept Test Plans*

Test Plan	Accept Times*				
IXC	$T_0 = 4.2$	$T_1 = 6.1$	$T_2 = 7.9$	$T_3 = 9.4$	$T_4 = 11.0$
	$T_5 = 12.4$	$T_6 = 13.9$	$T_7 = 15.3$	$T_8 = 16.6$	$T_9 = 18.0$
	$T_{10} = 19.3$	$T_{11} = 20.7$	$T_{12} = 22.0$	$T_{13} = 23.3$	$T_{14} = 24.5$
	$T_{15} = 25.8$	$T_{16} = 27.1$	$T_{17} = 28.3$	$T_{18} = 29.6$	$T_{19} = 30.8$
	$T_{20} = 32.1$	$T_{21} = 33.3$	$T_{22} = 34.5$	$T_{23} = 35.8$	$T_{24} = 37.0$
	$T_{25} = 38.2$	$T_{26} = 39.4$	$T_{27} = 40.6$	$T_{28} = 41.8$	$T_{29} = 43.0$
	$T_{30} = 44.2$	$T_{31} = 45.4$	$T_{32} = 46.6$	$T_{33} = 47.8$	$T_{34} = 49.0$
	$T_{35} = 50.1$	$T_{36} = 51.3$	$T_{37} = 52.5$	$T_{38} = 53.7$	$T_{39} = 54.8$
	$T_{40} = 56.0$	$T_{41} = 57.2$	$T_{42} = 58.3$	$T_{43} = 59.5$	$T_{44} = 60.7$
	$T_{45} = 61.8$	$T_{46} = 63.0$	$T_{47} = 64.1$	$T_{48} = 65.3$	$T_{49} = 66.5$
	$T_{50} = 67.6$	$T_{51} = 68.8$	$T_{52} = 69.9$	$T_{53} = 71.1$	$T_{54} = 72.2$
XC	$T_0 = 3.2$	$T_1 = 5.0$	$T_2 = 6.6$	$T_3 = 8.1$	$T_4 = 9.5$
	$T_5 = 10.9$	$T_6 = 12.2$	$T_7 = 13.6$	$T_8 = 14.9$	$T_9 = 16.1$
	$T_{10} = 17.4$	$T_{11} = 18.7$	$T_{12} = 19.9$	$T_{13} = 21.2$	$T_{14} = 22.4$
	$T_{15} = 23.6$	$T_{16} = 24.8$	$T_{17} = 26.1$	$T_{18} = 27.3$	$T_{19} = 28.4$
	$T_{20} = 29.6$	$T_{21} = 30.8$	$T_{22} = 32.0$	$T_{23} = 33.2$	$T_{24} = 34.4$
	$T_{25} = 35.6$	$T_{26} = 36.7$	$T_{27} = 37.9$	$T_{28} = 39.1$	$T_{29} = 40.2$
	$T_{30} = 41.4$	$T_{31} = 42.5$	$T_{32} = 43.7$	$T_{33} = 44.8$	$T_{34} = 46.0$
	$T_{35} = 47.1$	$T_{36} = 48.3$	$T_{37} = 49.4$	$T_{38} = 50.6$	$T_{39} = 51.7$
XIC	$T_0 = 3.0$	$T_1 = 4.8$	$T_2 = 6.3$	$T_3 = 7.8$	$T_4 = 9.2$
	$T_5 = 10.5$	$T_6 = 11.9$	$T_7 = 13.2$	$T_8 = 14.4$	$T_9 = 15.7$
	$T_{10} = 17.0$	$T_{11} = 18.2$	$T_{12} = 19.5$	$T_{13} = 20.7$	$T_{14} = 21.9$
	$T_{15} = 23.1$	$T_{16} = 24.3$	$T_{17} = 25.5$	$T_{18} = 26.7$	$T_{19} = 27.9$
	$T_{20} = 29.1$	$T_{21} = 30.3$	$T_{22} = 31.4$	$T_{23} = 32.6$	
XIIC	$T_0 = 3.7$	$T_1 = 5.6$	$T_2 = 7.2$	$T_3 = 8.8$	$T_4 = 10.3$
	$T_5 = 11.7$	$T_6 = 13.1$	$T_7 = 14.4$	$T_8 = 15.8$	$T_9 = 17.1$
	$T_{10} = 18.4$	$T_{11} = 19.7$	$T_{12} = 21.0$	$T_{13} = 22.3$	$T_{14} = 23.5$
	$T_{15} = 24.8$	$T_{16} = 26.0$			
XIIIC	$T_0 = 2.8$	$T_1 = 4.6$	$T_2 = 6.1$	$T_3 = 7.5$	$T_4 = 8.9$
	$T_5 = 10.3$	$T_6 = 11.6$	$T_7 = 12.9$	$T_8 = 14.1$	$T_9 = 15.4$
	$T_{10} = 16.6$	$T_{11} = 17.9$	$T_{12} = 19.1$		
XIVC	$T_0 = 2.7$	$T_1 = 4.4$	$T_2 = 5.9$	$T_3 = 7.3$	$T_4 = 8.7$
	$T_5 = 10.0$	$T_6 = 11.3$	$T_7 = 12.6$		
XVC	$T_0 = 3.5$	$T_1 = 5.4$	$T_2 = 7.0$	$T_3 = 8.6$	$T_4 = 10.0$
	$T_5 = 11.4$	$T_6 = 12.8$			
XVIC	$T_0 = 2.5$	$T_1 = 4.1$	$T_2 = 5.6$	$T_3 = 7.0$	$T_4 = 8.3$
XVIIC	$T_0 = 2.2$	$T_1 = 3.8$	$T_2 = 5.2$		
XIXC	$T_0 = 2.1$	$T_1 = 3.7$	$T_2 = 5.1$	$T_3 = 6.4$	$T_4 = 7.7$
	$T_5 = 8.9$	$T_6 = 10.2$	$T_7 = 11.4$	$T_8 = 12.6$	
XXC	$T_0 = 1.8$	$T_1 = 3.2$	$T_2 = 4.5$		
XXIC	$T_0 = 1.1$				

*Accept at time T_i if i failures have occurred to that time

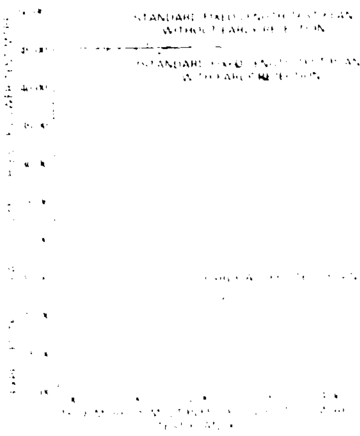


FIGURE 1

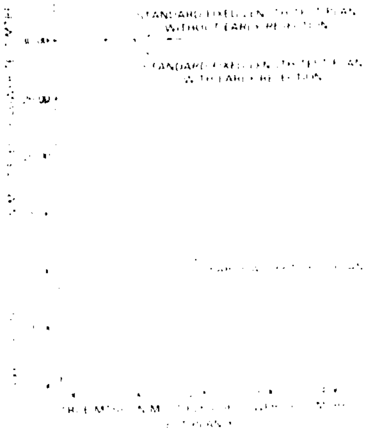


FIGURE 2

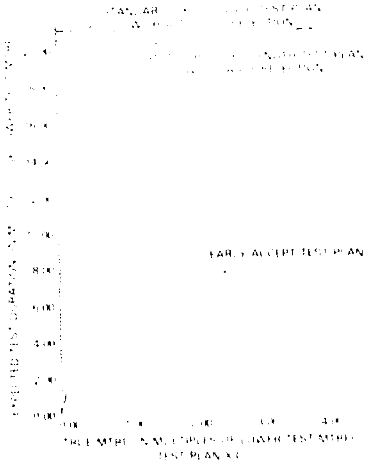


FIGURE 3

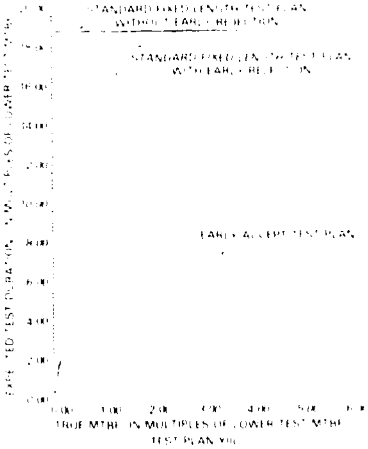


FIGURE 4

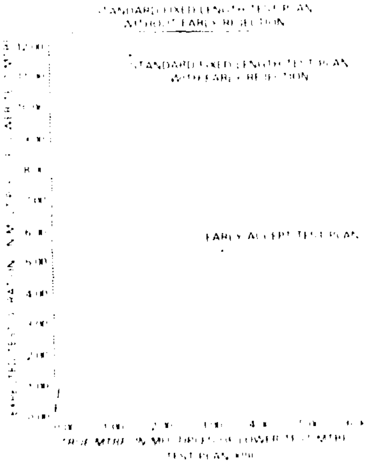


FIGURE 5

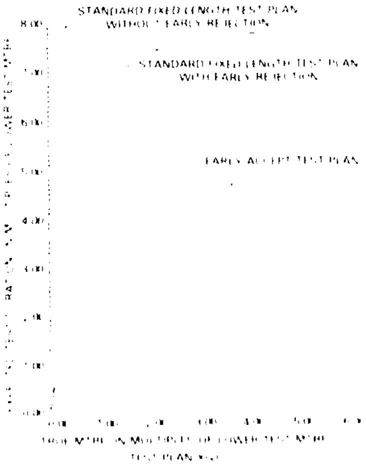


FIGURE 6

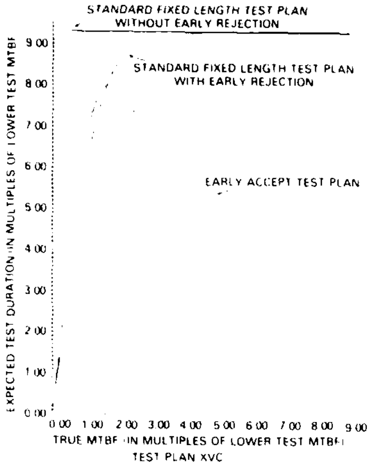


FIGURE 7

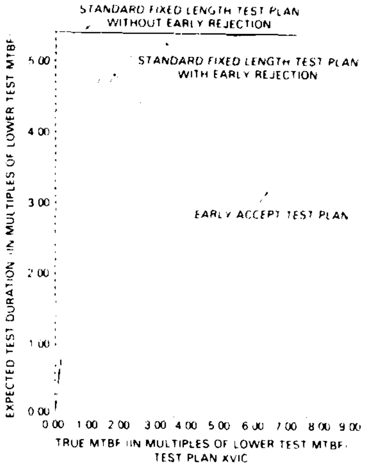


FIGURE 8

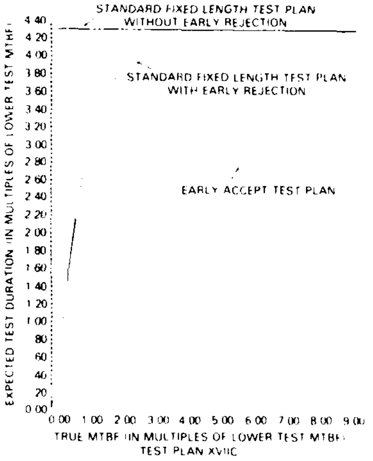


FIGURE 9

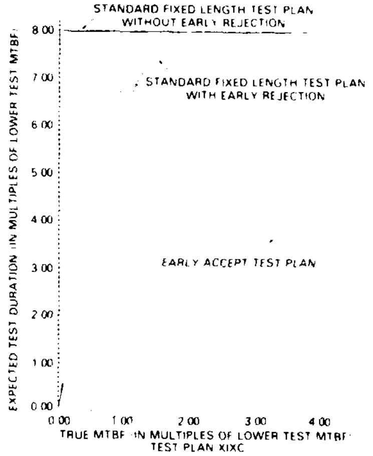


FIGURE 10

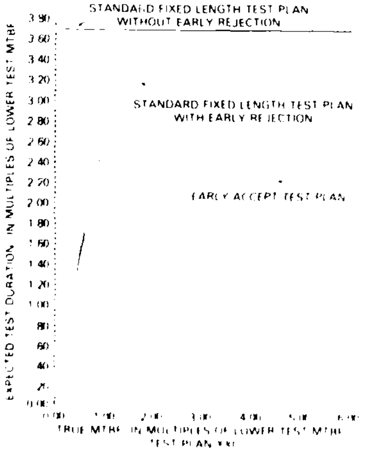
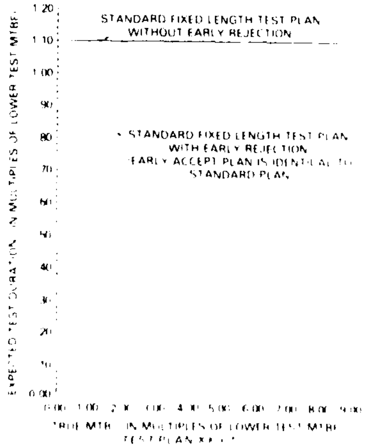


FIGURE 11



*THE STANDARD FIXED LENGTH TEST PLAN WITH EARLY REJECTION AND THE EARLY ACCEPT TEST PLAN ARE IDENTICAL FOR THIS CASE

FIGURE 12

cannot be conveniently used if an estimate of the true MTBF is required. If a standard test plan is used under these circumstances, the test continues even if a sufficient number of failures to reject occur prior to the total test *time*. A graph of this plan (without early rejection) also appears in the figures. It is not surprising that the early-accept test plans generally have smaller expected test durations.

The expected test durations are computed as follows. Let τ be the (random) test duration

(1)
$$E_\lambda[\tau] = E_\lambda\left[\tau \cdot \sum_{j} (I_{\{\text{accept at } T_j\}} + I_{\{\text{reject in } (T_{j-1}, T_j]\}})\right]$$
$$= \sum_{j} T_j \cdot A(j) + \sum_{j} E_\lambda\left[\tau \cdot I_{\{\text{reject in } (T_{j-1}, T_j]\}}\right],$$

where I_B denotes the indicator function of the event B, i.e., I_B equals 1 if the event occurs, 0 otherwise. To compute the terms in the second summation in (1), note that at least j and at most k failures must occur in $[0, T_{j-1}]$. (If fewer than j failures occur, the test will accept by time T_{j-1}; if more than k failures occur, the test will reject by time T_{j-1}.) Given that r failures occur in $[0, T_{j-1}]$ $(j \leqslant r \leqslant k)$ and given that the test does not terminate by time T_{j-1}, the test will reject in $(T_{j-1}, T_j]$ if and only if $k+1-r$ failures occur in $(T_{j-1}, T_j]$. By the memoryless property of the exponential distribution, the expected test time under these conditions is

$$\int_0^{T_j - T_{j-1}} (T_{j-1} + z) f(z)\, dz$$

where $f(z)$ is a gamma density with parameters λ and $k+1-r$. Thus, by a conditional expectation argument

(2)
$$E_\lambda[\tau \cdot I_{\{\text{reject in } (T_{j-1}, T_j]\}}] = \sum_{r=j}^{k} Q(j-1, r) \cdot \int_0^{T_j - T_{j-1}} (T_{j-1} + z) f(z)\, dz\,,$$

where $Q(j,r) = Pr$ {do not accept or reject at or before time T_j, and r failures in $[0, T_j]$}, $(j \leqslant r)$.

THEOREM 2: For $j \leqslant r \leqslant k$

$$Q(j,r) = \frac{(\lambda T_j)^r \exp(-\lambda T_j)}{r!} - \sum_{l=0}^{j} A(l) \cdot \frac{[\lambda(T_j - T_l)]^{r-l} \exp(-\lambda(T_j - T_l))}{(r-l)!}$$

PROOF: For $j \leqslant r \leqslant k$,

$$\Pr\{\text{exactly } r \text{ failures in } [0, T_j]\} = Pr\left\{\bigcup_{l=0}^{j} \{\text{accept at time } T_l \text{ (necessarily with } l \text{ failures) and } (r-l) \text{ failures in } (T_l, T_j]\}\right\} + Q(j,r)$$

$$\frac{(\lambda T_j)^r \exp(-\lambda T_j)}{r!} = \sum_{l=0}^{j}\left[A(l) \cdot \frac{[\lambda(T_j - T_l)]^{r-l} \exp(-\lambda(T_j - T_l))}{(r-l)!}\right] + Q(j,r).$$

All that remains is to compute the integral in (2). This integral can be expressed in terms of the incomplete gamma distribution and evaluated by standard computer subroutines [1].

5. CONCLUSIONS

From an operating characteristic curve point of view it makes no difference whether a fixed-length test plan, a probability ratio sequential test plan (truncated) or an early-accept plan is chosen, provided each is designed to have the same operating characteristics. Generally, the ordering of the expected test duration is smallest for the probability ratio sequential test plan, followed by the early-accept plan, and largest for the fixed-length plan. The advantage of the early-accept plan, over the probability ratio sequential test plan, is purely psychological. The producer never has a lot rejected early, and early decisions occur only with the desirable outcome (from the producer's point of view) of acceptance. Such an advantage cannot be discounted.

ACKNOWLEDGMENT

The authors are indebted to Vlad Rutenberg for help in carrying out some of the calculations.

REFERENCES

[1] *IMSL Library Reference Manual*, Volume 2, International Mathematical and Statistical Libraries, Inc., Houston, Texas, 7th Edition, February 1979.

[2] *Military Standard 781C, Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution*, U.S. Department of Defense, AMSC Number 22333, 21 October 1977.

A TWO-STATE SYSTEM WITH PARTIAL AVAILABILITY IN THE FAILED STATE

Laurence A. Baxter*

University of Delaware
Newark, Delaware

ABSTRACT

A generalization of the alternating renewal model of a repairable system to permit partial availability in the failed state is introduced. It is shown how, by making use of an embedded alternating renewal process, we can readily derive expressions for various measures of system availability. Expressions for the point availability of the generalized process are presented.

1. INTRODUCTION

Consider a two-state system, i.e., a machine subject to stochastic failure and repair. If it is assumed that the sequences of periods of operation and repair constitute an alternating renewal process, a variety of expressions for predicting the availability of the system, known as *availability measures*, may be derived (see, for example, Baxter [2]). These formulae can readily be evaluated by means of the cubic splining algorithm of Cléroux and McConalogue [4] (see also McConalogue [5], [6]).

The model assumes that a breakdown will wholly incapacitate the system, but this need not be the case, e.g., a large machine dependent on auxiliaries may be able to operate at a reduced capacity if some of the auxiliaries fail. An example of such a machine is a coal-fired boiler in which the fuel is supplied by a number of mills: while the failure of one or more of the mills will reduce the effectiveness of ,he boiler, a total breakdown will not necessarily occur. In this paper we present a generalization of the two-state system which permits partial availability in the failed state. It will be shown that we can formulate this generalized model in terms of an embedded alternating renewal process and hence make use of existing theory and numerical techniques.

It is first necessary to introduce some notation. Let F and G denote the distribution functions of the failure and repair times respectively and suppose that these have finite expectations and variances μ_1, μ_2, σ_1^2, and σ_2^2, respectively. Define the indicator variable of the two-state system

$$I_k(t) = \begin{cases} 1 & \text{if the system is operating at } t \\ 0 & \text{otherwise} \end{cases}$$

*This research was performed while the author was at University College, London, England.

where $k = 0(1)$ if the system enters the down (up) state at $t = 0$. We define the Stieltjes convolution of two functions, P and Q say, each with support on the nonnegative real line as

$$P * Q(t) = \int_0^t P(t-u)\, dQ(u)$$

and the n-fold recursive convolution of $P(t)$ is denoted $P^{(n)}(t)$. The point availability of the two-state system is defined as $A_k(t) = p\{I_k(t) = 1\}$. It can be shown that

(1) $$A_1(t) = \bar{F}(t) + \bar{F} * H(t)$$

(2) $$A_0(t) = G(t) - \bar{G} * H(t)$$

where

(3) $$H(t) = \sum_{n=1}^{\infty} F^{(n)} * G^{(n)}(t)$$

denotes the renewal function of the sequence of failures (repairs) embedded in [illegible] renewal process if there is a failure (repair) at $t = 0$ and where $\bar{F}(t) = 1 - F(t)$ [illegible] function $P(t)$ such that $0 \leqslant P(t) \leqslant 1$ for all t (see, for example, Baxter [2]).

2. THE GENERALIZED MODEL

There are many ways of generalizing the alternating renewal model to allow for partial availability in the failed state. We could, for example, assume n levels of partial availability and hence generalize the two-state system to an $(n+1)$-state semi-Markov process. This would result in a considerably more complex model for a relatively little increase in generality.

The approach adopted here is to assume that a proportion γ, $(0 \leqslant \gamma \leqslant 1)$, of breakdowns exhibit partial availability and that the level, Λ, is a random variable, independent of the failure and repair times, with distribution function M. The value of Λ is assumed to remain constant during any given period of repair. The distribution M is conditional on $\{\Lambda > 0\}$ (although we could equally consider a distribution which assigns a mass of probability $1 - \gamma$ to the value 0). This model is equivalent to a three-state semi-Markov process with transition matrix

$$\begin{array}{cl} \begin{pmatrix} 0 & \gamma & 1-\gamma \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} & \begin{array}{l} \text{Available} \\ \text{Partially available} \\ \text{Wholly unavailable.} \end{array} \end{array}$$

We now define the *multistate variable*

$$J(t) = \begin{cases} 1 & \text{if the system is fully available at } t \\ \lambda & \text{if the system is only operating at level } \lambda \ (0 < \lambda < 1) \text{ at } t \\ 0 & \text{if the system is wholly unavailable at } t. \end{cases}$$

In particular, $\{J_k(t),\ t > 0\}$ denotes the generalized process in which there is a failure (repair) at $t = 0$ if $k = 0(1)$.

A variety of types of availability measure can be defined for the process $\{J(t),\ t > 0\}$. We could, for example, consider the expectation; in particular

(4) $$E\{J_1(t)\} = A_1(t) + \gamma\, E(\Lambda)\, \bar{A}_1(t)$$

(5) $$E\{J_0(t)\} = A_0(t) + \gamma\, E(\Lambda)\, \bar{A}_0(t).$$

Similarly, the expected proportion of time for which the system is wholly or partially available in $(0,t]$ is given by

$$E\left\{\frac{1}{t}\int_0^t J_1(u)\,du\right\} = a_1(t) + \gamma E(\Lambda)\,\bar{a}_1(t) \tag{6}$$

$$E\left\{\frac{1}{t}\int_0^t J_0(u)\,du\right\} = a_0(t) + \gamma E(\Lambda)\,\bar{a}_0(t) \tag{7}$$

where $a_i(t) = \frac{1}{t}\int_0^t A_i(u)\,du$ denotes the average availability of the process $\{I_i(t),\ t > 0\}$ ([1], p. [illegible]).

EXAMPLE

Consider the alternating Poisson process, i.e., $F(t) = 1 - e^{-\nu t}$ and $G(t) = 1 - e^{-\mu t}$, in this case it is well known that

$$A_1(t) = \frac{\mu}{\nu + \mu} + \frac{\nu}{\nu + \mu}\,e^{-(\nu+\mu)t}$$

$$A_0(t) = \frac{\mu}{\nu + \mu} - \frac{\mu}{\nu + \mu}\,e^{-(\nu+\mu)t}.$$

Suppose that $\Lambda \sim$ Beta (α, β) and hence $E(\Lambda) = \alpha/(\alpha + \beta)$. Substituting these expressions into (4) and (5) gives us the following formulae for $E\{J_1(t)\}$ and $E\{J_0(t)\}$.

$$E\{J_1(t)\} = \frac{\mu\alpha + \mu\beta + \alpha\gamma\nu}{(\nu + \mu)(\alpha + \beta)} + \frac{\nu(\alpha + \beta - \alpha\gamma)}{(\nu + \mu)(\alpha + \beta)}\,e^{-(\nu+\mu)t}$$

$$E\{J_0(t)\} = \frac{\mu\alpha + \mu\beta + \alpha\gamma\nu}{(\nu + \mu)(\alpha + \beta)} - \frac{\mu(\alpha + \beta - \alpha\gamma)}{(\nu + \mu)(\alpha + \beta)}\,e^{-(\nu+\mu)t}.$$

3. THE AUGMENTED PROCESS

The expectation of the multistate variaole is of limited use as there is no obvious extension to, for example, interval availability, i.e., $p\{I(t) = 1\ \forall\ t \in [t_1, t_2]\}$ [2]. Further, this measure is not very sensitive to the distribution of Λ; only $E(\Lambda)$ is required and hence identical forecasts would result from two distributions with the same mean but different variances. Observe also that, for $\gamma > 0$, $E\{J(t)\} > A(t)$ and hence $E\left\{\frac{1}{t}\int_0^t J(u)\,du\right\} > a(t)$. Thus, any positive value of Λ, no matter how small, increases the measure of system availability. This could be an unrealistic assumption in practice: if Λ is close to 0, it may not be worthwhile attempting to use the machine until it is fully repaired.

An alternative approach, which is more likely to be of use in practice, is to regard the system as operating satisfactorily if $\Lambda \geq \lambda_0$ and as broken down otherwise. The system can thus undergo an arbitrary number of changes of state without becoming unavailable provided that each repair period exhibits partial availability at a level exceeding λ_0. The alternating sequence of periods of availability at a level no less than λ_0 and periods of repair in which $\Lambda < \lambda_0$ clearly constitutes an embedded alternating renewal process for fixed λ_0. Thus, if we define

$$\tilde{I}(t) = \begin{cases} 1 & \text{if } J(t) \geq \lambda_0 \\ 0 & \text{otherwise} \end{cases}$$

we can apply the arguments of Baxter [2] to derive expressions for probabilities of the form $p\{J(t) > \lambda_0 \; \forall \; t \in T\}$, where T is an index set comprising an arbitrary (finite) series of points and intervals, for the process $\{\tilde{I}(t),\ t > 0\}$, which we call the *augmented process*. In particular, we shall consider $\{\tilde{I}_k(t),\ t > 0\}$, the augmented process in which there is a failure with $\Lambda < \lambda_0$ (repair) at $t = 0$ for $k = 0(1)$. It is important to appreciate that the interpretation of the subscript k is not the same for functions defined with respect to the two-state system and those defined with respect to the augmented process. For the former, the values 0 and 1 are used to denote a failure and repair at $t = 0$ respectively, whereas for the latter, these values denote a failure at $t = 0$ such that the level of partial availability during the succeeding downtime is less than λ_0 and a repair at $t = 0$, respectively. If $\gamma = 0$, the augmented process degenerates to the two-state system and the interpretations of the two subscripts coincide.

Let Ξ denote the duration of an "uptime" in the augmented process, i.e., the time from a repair following a downtime with $\Lambda < \lambda_0$ to the beginning of the next such downtime, and suppose that this has distribution function Φ. It is easily seen that

$$(8) \qquad \Phi(t) = (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n F^{(n+1)} * G^{(n)}(t)$$

where $\alpha = \gamma p\{\Lambda > \lambda_0\} = \gamma \bar{M}(\lambda_0)$ and where

$$G^{(0)}(t) = \begin{cases} 1 & \text{if } t \geqslant 0 \\ 0 & \text{if } t < 0. \end{cases}$$

We can readily derive expressions for the mean and variance of Ξ by means of conditional expectation:

$$(9) \qquad E(\Xi) = \frac{\mu_1 + \alpha\mu_2}{1 - \alpha}$$

$$(10) \qquad \operatorname{var}(\Xi) = \frac{\alpha}{1 - \alpha}(\sigma_1^2 + \sigma_2^2) + \frac{\alpha}{(1 - \alpha)^2}(\mu_1 + \mu_2)^2 + \sigma_1^2.$$

Observe that if $\alpha = 1$, both mean and variance are infinite. This is to be expected as in this case the system is always available. Similarly, if $\alpha = 0$, the augmented process reduces to the alternating renewal model and $E(\Xi) = \mu_1$, var $(\Xi) = \sigma_1^2$.

EXAMPLE

Consider the alternating Poisson process. The Laplace-Stieltjes transforms of F and G are given by $f^*(s) = \nu/(s + \nu)$ and $g^*(s) = \mu/(s + \mu)$ respectively, and hence the Laplace-Stieltjes transform of Φ is

$$\phi^*(s) = \frac{\nu(1 - \alpha)(s + \mu)}{(s + \nu)(s + \mu) - \alpha\nu\mu}$$

$$= \nu(1 - \alpha)\left[\frac{s}{(s + A)(s + B)} + \frac{\mu}{(s + A)(s + B)}\right]$$

where $\qquad A,B = \frac{1}{2}\left[-(\nu + \mu) \pm \sqrt{\{(\nu + \mu)^2 - 4\nu\mu(1 - \alpha)\}}\right].$

Thus, on inversion, we see that the density of Ξ is

$$\phi(t) = \frac{\nu(1 - \alpha)}{A - B}\left[Ae^{At} - Be^{Bt} - \mu(e^{At} - e^{Bt})\right].$$

Observe that $\phi(t)$ is a special case of the density of the first passage time to absorption in the Chiang-Hsu alternating renewal process with an absorbing state [3].

4. POINT AVAILABILITY

The point availability $\tilde{A}(t) = p\{\tilde{I}(t) = 1\}$ of the augmented process is the probability that the system is available at t or that it is under repair and that the level of partial availability exceeds λ_0. The following expressions for $\tilde{A}_1(t)$ and $\tilde{A}_0(t)$ are obtained by substituting

$$\phi^*(s) = \frac{(1-\alpha)f^*(s)}{1-\alpha f^*(s)g^*(s)}$$

and $g^*(s)$ into the formulae for $\tilde{A}_1^*(s)$ and $\tilde{A}_0^*(s)$, performing some rearrangement and inverting:

(11) $$\tilde{A}_1(t) = A_1(t) + \alpha \bar{A}_1(t)$$

(12) $$\tilde{A}_0(t) = (1-\alpha)A_0(t) + \alpha G(t).$$

As would be expected, $\tilde{A}_k(t) = A_k(t)$ if $\alpha = 0$ $(k = 0, 1)$ as in this case $\{\tilde{I}_k(t), t > 0\}$ reduces to $\{I_k(t), t > 0\}$. Similarly, if $\alpha = 1$ the system cannot fail and hence $\tilde{A}_1(t) = 1$ and $\tilde{A}_0(t) = G(t)$.

Expression (11) clearly corresponds to (4) whereas expression (12) does not correspond so obviously to (5); an interpretation of this result is, however, more evident if we make use of (2) to rewrite (12) as

(13) $$\tilde{A}_0(t) = A_0(t) + \alpha\bar{G} * H(t).$$

We now see that we are increasing $A_0(t)$, the point availability of the two-state system, by the probability that the system fails at $u < t$ and that the succeeding repair time, which exhibits partial availability at a level exceeding λ_0, is greater than $t - u$, for each $u \in (0,t]$.

EXAMPLE

On substituting the formulae for the point availabilities of the alternating Poisson process into (11) and (12) we obtain the following expressions for the point availabilities of the corresponding augmented process:

$$\tilde{A}_1(t) = \frac{\mu + \alpha\nu}{\nu + \mu} + \frac{(1-\alpha)\nu}{\nu + \mu}e^{-(\nu+\mu)t}$$

$$\tilde{A}_0(t) = \frac{\mu + \alpha\nu}{\nu + \mu} - \frac{(1-\alpha)\mu}{\nu + \mu}e^{-(\nu+\mu)t} - \alpha e^{-\mu t}.$$

On applying the key renewal theorem to (4), (5), (11) and (12), we see that

$$\lim_{t\to\infty} \tilde{A}(t) = \frac{\mu_1 + \alpha\mu_2}{\mu_1 + \mu_2}$$

$$\text{cf} \lim_{t\to\infty} E\{J(t)\} = \frac{\mu_1 + \gamma E(\Lambda)\mu_2}{\mu_1 + \mu_2}.$$

Expressions for other availability measures are readily derived but, in general, we do not obtain formulae which, like those for $\tilde{A}_k(t)$, are simple modifications of the corresponding expressions for the alternating renewal model.

ACKNOWLEDGMENTS

I am indebted to Mr. R. F. Galbraith for several helpful discussions during the preparation of this paper. I would also like to thank Mr. M. A. Baxter for some useful comments and the referee for a number of suggestions which improved the presentation.

REFERENCES

[1] Barlow, R.E. and F. Proschan, *Statistical Theory of Reliability and Life Testing* (Holt, Rinehart and Winston, New York, 1975).

[2] Baxter, L.A., "Availability Measures for a Two-State System," Journal of Applied Probability *18* (1981) (to appear).

[3] Chiang, C.L. and J.P. Hsu, "An Alternating Renewal Process with an Absorbing State" in *Applications of Statistics*, 109-121, Editor, P.R. Krishnaiah (North Holland, Amsterdam, 1977).

[4] Cléroux, R. and D.J. McConalogue, "A Numerical Algorithm for Recursively-Defined Convolution Integrals Involving Distribution Functions," Management Science *22*, 1138-1146 (1976).

[5] McConalogue, D.J., "Convolution Integrals Involving Probability Distribution Functions" (Algorithm 102), Computer Journal *21*, 270-272 (1978).

[6] McConalogue, D.J., "Numerical Treatment of Convolution Integrals Involving Distributions with Densities having Singularities at the Origin," Communications in Statistics, B (1981) (to appear).

AN ANALYSIS OF SINGLE ITEM INVENTORY SYSTEMS WITH RETURNS*

John A. Muckstadt and Michael H. Isaac

Cornell University
Ithaca, New York

ABSTRACT

Inventory systems with returns are systems in which there are units returned in a repairable state, as well as demands for units in a serviceable state, where the return and demand processes are independent. We begin by examining the control of a single item at a single location in which the stationary return rate is less than the stationary demand rate. This necessitates an occasional procurement of units from an outside source. We present a cost model of this system, which we assume is managed under a continuous review procurement policy, and develop a solution method for finding the policy parameter values. The key to the analysis is the use of a normally distributed random variable to approximate the steady-state distribution of net inventory.

Next, we study a single item, two echelon system in which a warehouse (the upper echelon) supports $N (N \geqslant 1)$ retailers (the lower echelon). In this case, customers return units in a repairable state as well as demand units in a serviceable state at the retailer level only. We assume the constant system return rate is less than the constant system demand rate so that a procurement is required at certain times from an outside supplier. We develop a cost model of this two echelon system assuming that each location follows a continuous review procurement policy. We also present an algorithm for finding the policy parameter values at each loca ion that is based on the method used to solve the single location problem.

1. INTRODUCTION

Many models have been developed during the past 15 years pertaining to various aspects of managing repairable item inventory systems (e.g., [1],[4],[10],[11],[12],[15], and [16]).[1] Most of these models contain the assumption that the failure of a unit simultaneously generates a demand for a unit of exactly the same type, i.e., the demand process for serviceable units and the return processes of failed units are perfectly correlated.

In certain instances, however, this assumption of perfect correlation between the demand and return processes is not valid. For example, this can occur in situations where equipment is leased, rented, and/or sold, such as found in the telephone, computer and copying machine industries. Returns do not necessarily correspond to failures in these cases, but rather to lease or rental expirations. At the time a unit is returned, it may have to go through a repair or

*This research was supported in part by the Office of Naval Research under Contract N00014-75-C-1172

[1] A repairable item is an item which fails, but which can be repaired and subsequently made available to satisfy a future demand or an existing backorder

overhaul process before reissue. There is no reason to assume that the customer will request a unit of exactly the same type when a lease or rental agreement expires. Similarly, when a customer requests a particular type of unit, there is no reason to assume that the customer will return one of exactly the same type.

The authors studied a real two echelon inventory repair system managed by a manufacturer of reprographic equipment. This system closely resembles the one described in Section 3. For that system we found the demand and return processes to be independent Poisson processes. That is, we tested and could not reject the hypotheses that the demand and return random variables had Poisson distributions, and that the return and demand random variables were independent. The research described in this paper reflects our study of this system's behavior. Consequently, we assume in the remainder of this paper that the demand and return processes are independent. We call such inventory systems, inventory systems with returns.

Only a few papers have been published on inventory systems with returns. These papers contain simplifying assumptions which make them of limited practical value. Heyman [6,7] considers optimal disposal policies for a single item inventory system with returns; but his assumptions include instantaneous outside procurement (implying no backorders or lost sales) and no fixed cost of ordering (implying no lot size reordering). Hoadley and Heyman [8] consider a two echelon inventory system with outside procurement, returns, disposals, and transshipment; but their model is a one period model, and all of the mentioned transactions are assumed to occur instantaneously. Simpson [16] develops the optimal solution for a finite horizon, periodic review model. His model allows for correlation between the return and demand processes. Backlogging is permitted, but both repairs and outside procurements are assumed to be instantaneous.

For the most part, the methods of analysis in these three papers rely heavily upon the assumptions of instantaneous repair and procurement. Their approaches are of little use when analyzing situations in which repair and procurement times are not zero.

Finally, Schrady [14] solves for repair carcass and procurement lot sizes for a completely deterministic system. Gajdalo [2] extends this to a 'continuous review repair policy' for an inventory system with stochastic (compound Poisson) returns and demands. He uses computer simulation to test several heuristics for computing the reorder point and lot sizes for both procuring and repairing items. All lead times, including repair times, are assumed constant.

Our approach differs substantially from those taken in these previous studies. We begin in the next section by analyzing a single item, single location inventory system with returns. We develop the stationary distribution of two key random variables that describe the probabilistic behavior of the inventory system. This analysis is used as the basis for a cost model. A solution method is then presented for finding the values of the policy parameters. The results of the single echelon case are then extended in Section 3 to a specific two echelon situation, which corresponds to the real environment mentioned earlier. In Section 4, we conclude with a brief summary and some final comments.

2. THE SINGLE ECHELON CASE

The system we study in this section consists of a single type of item managed at a single location. A schematic representation of the system's operation is given by Figure 1. As shown, this location is assumed to contain both a repair facility for returned units and a warehouse, or storage facility, for serviceable inventory.

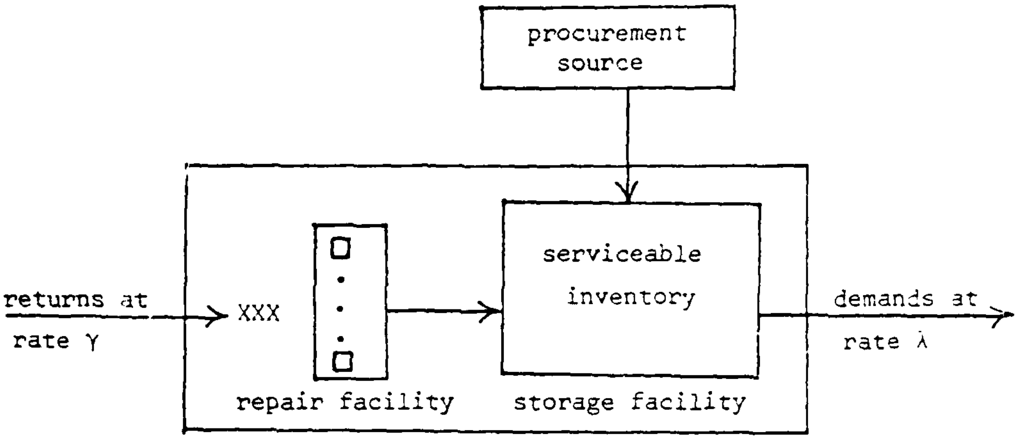


FIGURE 1. A schematic representation of the inventory system

We assume returns of repairable units occur as a Poisson process with rate γ, and demands for serviceable units occur as a Poisson process with rate λ. As we have stated, we also assume that these two processes are independent. γ is assumed to be less than λ, so that an occasional procurement of units from an outside source is required. Units procured in this manner arrive in a serviceable state τ time units after they are ordered.

The repair facility behaves as a first-come, first-served queueing system with Poisson arrivals (the Poisson returns). All returned units require repair, and repair times of returned units are independent. Since $\gamma < \lambda$, the repair system is always operating as long as repairables are present. No other assumptions about the queueing repair system (e.g., service time distribution or the number of repair servers) are made.

The output of this queueing repair system is input to the stock of on-hand serviceable inventory, as is the arrival of outside procurement orders.

All demands that are not satisfied immediately are assumed to be backordered.

We define 'net inventory' at a point in time to be the number of on-hand serviceable units in the storage facility minus the number of outstanding backorders. We also define 'inventory position' at a point in time to be the sum of net inventory, the number of units in the repair queueing system, and the number of units on order from the outside procurement source.

Let

$I(t)$ = the inventory position at time t,

$N(t)$ = the net inventory at time t,

$R(t)$ = the number of units in the repair queueing system at time t,

$P(t)$ = the number of units on order from the outside supplier at time t,

$O(t)$ = the on-hand serviceable inventory at time t,

and $B(t)$ = the number of outstanding backorders at time t.

Then

$$I(t) = N(t) + R(t) + P(t),$$

and

$$N(t) = O(t) - B(t).$$

Our final assumption concerns the form of the procurement policy. We assume that a continuous review (Q,r) procurement policy is followed, i.e., when the inventory position drops below $r + 1$, and order for $Q \geqslant 1$ units is placed immediately. Since the repair queueing system is assumed to be operating continuously, our objective is simply to find values of Q and r

Our analysis begins with the derivation of the steady-state distribution of inventory position. This result is used in the derivation of an approximation of the steady-state distribution of net inventory, and is followed by a discussion of the accuracy of the approximation.

2.1 Derivation of the Stationary Distribution of Inventory Position

Changes in the state of the inventory position are caused only by demands and returns. State $i (i = r + 1, r + 2, \ldots)$ can be entered from state $i + 1$ when a demand for a serviceable item occurs; state j $(j = r + 2, r + 3, \ldots)$ can be entered from state $j - 1$ when an item is returned. In addition, state $r + Q$ can also be reached from state $r + 1$ when a serviceable item is demanded (an order for Q units is placed immediately when the inventory position drops below $r + 1$). The time between state transitions is exponentially distributed, since the return and demand processes are Poisson processes. The state transition flow diagram is given in Figure 2, with the transition rates as indicated.

Let $u_i = \lim_{t \to \infty} \text{Prob}(I(t) = r + 1 + i)$, the stationary probability that inventory position is equal to $r + 1 + i$. This limit exists because the states of this system are the states of an irreducible, ergodic, Markov chain [13]. The steady-state balance equations corresponding to this system are

(1)
$$(\lambda + \gamma)u_0 = \quad + \lambda u_1,$$
$$(\lambda + \gamma)u_1 = \gamma u_0 + \lambda u_2,$$
$$(\lambda + \gamma)u_2 = \gamma u_1 + \lambda u_3,$$

$$(\lambda + \gamma)u_{Q-1} = \gamma u_{Q-2} + \lambda u_Q + \lambda u_0$$
$$(\lambda + \gamma)u_Q = \gamma u_{Q-1} + \lambda u_{Q+1},$$

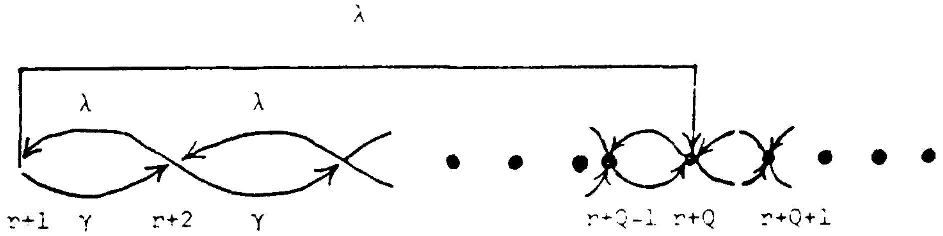


FIGURE 2. State transition flow diagram for inventory position

A generating function approach can be used to solve for the u_i. Define the generating function $G(z)$ to be $G(z) = \sum_{i=0}^{\infty} z^i u_i$. Using (1) we find that

$$(2) \qquad G(z) = \frac{\lambda - \gamma}{Q} \frac{(1 - z^Q)}{(1 - z)(\lambda - \gamma z)},$$

from which we find that the u_i are given by

$$(3) \qquad u_i = \begin{cases} 0 & i < 0, \\ \dfrac{1 - \left(\dfrac{\gamma}{\lambda}\right)^{i+1}}{Q}, & 0 \leqslant i \leqslant Q - 1, \\ \dfrac{\left(\dfrac{\gamma}{\lambda}\right)^{i-Q+1}\left[1 - \left(\dfrac{\gamma}{\lambda}\right)^{Q}\right]}{Q}, & i \geqslant Q, \end{cases}$$

and the mean and variance of the stationary distribution of inventory position are given by

$$(4) \qquad E[\lim_{t\to\infty} I(t)] = r + 1 + G'(1) = r + 1 + \frac{Q-1}{2} + \frac{\gamma}{\lambda - \gamma},$$

and

$$(5) \qquad \mathrm{Var}[\lim_{t\to\infty} I(t)] = G''(1) + G'(1) - [G'(1)]^2 = \frac{Q^2 - 1}{12} + \frac{\lambda\gamma}{(\lambda - \gamma)^2},$$

respectively.

If $Q = 1$, Figure 2 is the transition flow diagram for an $M/M/1$ queueing system in which the 'arrival' rate is γ and the 'service' rate is λ. In this case (3) reduces to the geometric distribution, which is the steady-state distribution of the number of customers present in an $M/M/1$ system.

Note that when $\gamma = 0$, (4), (5), and (3) reduce to the mean, variance, and probability distribution, respectively, of a uniformly distributed random variable, which is a well known result (see Reference 5).

2.2 An Approximation to the Stationary Distribution for Net Inventory

Next, we develop an approximation to the stationary distribution of net inventory, which is the basis for the cost model used to determine optimal values of Q and r.

Recall that τ, the procurement lead time, is constant. Thus, any units on order at time $t - \tau$ will have arrived by time t. Similarly, any order placed after time $t - \tau$ will not have arrived by time t. Therefore, we see that

$$(6) \qquad N(t) = I(t - \tau) - R(t - \tau) + Z(t - \tau,t) - D(t - \tau,t),$$

where $R(t - \tau)$ = the number of units in the repair system at time $t - \tau$,

$Z(t - \tau,t)$ = the output of the repair system in the interval $(t - \tau,t]$,

and $D(t - \tau,t)$= the number of demands in the interval $(t - \tau,t]$.

$R(t - \tau)$ is subtracted from $I(t - \tau)$ so that we do not double count the units in the repair system at time $t - \tau$ that complete service by time t. Therefore, net inventory at time t consists of units on order, already serviceable, or backordered at time $t - \tau$ (all measured in $I(t - \tau)$), plus those units completing repair by time $t - \tau$, minus demands over the interval $(t - \tau,t]$.

Let us separately examine the individual terms of (6). The steady-state distribution of $I(t - \tau)$ has already been obtained. The number of demands over the interval $(t - \tau,t]$ is Poisson distributed with mean $\gamma\tau$ and is independent of the other three random variables on the right-hand side of Equation (6).

The distributions of $R(t - \tau)$ and $Z(t - \tau,t)$ are readily available for many queueing systems; but, they are not independent of each other or of $I(t - \tau)$. The joint distribution of these random variables is difficult to develop analytically. Consequently, an approximation to the distribution of net inventory will be developed, using (6), rather than developing the exact distribution.

We initially observed that the steady-state distribution of net inventory for numerous test cases (obtained via simulation) closely resembled a normal distribution. As a result, the normal distribution was considered to be a candidate approximation to the steady-state distribution of net inventory.

Equation (6) is used to determine the mean μ and to approximate the variance, σ^2, of this normal approximation. Letting $t \rightarrow \infty$, we have

$$(7) \qquad \mu = E(N(t)) = E(I(t - \tau)) - E(R(t - \tau)) + E(Z(t - \tau,t)) - E(D(t - \tau,t))$$
$$= r + 1 + \frac{Q - 1}{2} + \frac{\gamma}{\lambda - \gamma} - E(R(t - \tau)) + \gamma\tau - \lambda\tau,$$

using (5), and noting that the expected output of a queueing system over an interval is equal to the expected input over an interval of the same length. Also, by ignoring covariance terms, we approximate σ^2 by

$$(8) \qquad \sigma^2 = \text{Var}(N(t)) \approx \text{Var}(I(t - \tau)) + \text{Var}(R(t - \tau))$$
$$+ \text{Var}(Z(t - \tau,t)) + \text{Var}(D(t - \tau,t))$$
$$= \frac{Q^2 - 1}{12} + \frac{\lambda\gamma}{(\lambda - \gamma)^2} + \text{Var}(R(t - \tau)) + \text{Var}(Z(t - \tau,t)) + \lambda\tau,$$

using (5). Note that exact expressions and good approximations for $E(R(t - \tau))$, $\text{Var}(R(t - \tau))$, and $\text{Var}(Z(t - \tau,t))$ are available for many queueing systems (e.g., see [3]).

The accuracy of the normal approximation, whose mean and variance are given by (7) and (8), was tested using an incomplete factorial experiment. The variable factors were the number of repair servers, the repair service distribution, the repair system traffic intensity, the procurement lead time τ, the procurement lot size Q, and the ratio γ/λ. In each test case, the accuracy of the normal approximation was first measured by finding the area between the normal curve and the curve representing the continuous version of the distribution of net inventory, which was obtained via simulation.

The conclusion drawn from this experiment was that the major factor affecting the accuracy of the normal approximation is the ratio of the return rate to the demand rate, γ/λ. In fact, the normal approximation is quite accurate when $\gamma/\lambda < .6$. However, a closer analysis of the normal curves revealed that the normal approximation was an excellent one in the left-hand tail of the distribution of net inventory in *all* the test cases. (We discuss in Section 2.3 why the left-hand tail of the distribution is all that is needed to determine optimal values for Q and r.) The difference between left-hand tail percentiles of the experimental distributions and the corresponding approximating normal distributions were computed. The percentiles never differed by more than a few percent. Based on this observation we conclude that the steady-state distribution of net inventory can be accurately approximated by a normal distribution whose mean and variance are given by (7) and (8), respectively.

2.3 Cost Model and Solution Method

The optimization model we will construct includes a fixed procurement order cost, a holding cost, and a time-weighted backorder cost. In particular, let

A = the fixed procurement order cost ($/procurement order),

h = the holding cost ($/unit-year),

and $\hat{\pi}$ = the backorder cost ($/unit-year).

Our objective function, K, is the sum of the expected annual procurement ordering, holding, and backorder costs. It will be evaluated by taking the sum of

(1) A × (the expected number of orders placed per year),

(2) h × (the expected serviceable on-hand inventory at a random point in time),

and (3) $\hat{\pi}$ × (the expected number of outstanding backorders at a random point in time).

Both the expected on-hand inventory and expected backorders at a random point in time will be calculated using the normal approximation to the distribution of net inventory.

Note that we need not consider holding costs charged against units in repair. Due to the assumption that no inserted idleness in the queueing repair system is allowed, these holding costs are independent of the values of the procurement policy parameters.

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the standard normal density and standard normal distribution functions, respectively. Let $h(x)$ be the normal density, which is the continuous approximation to the steady-state distribution of net inventory, whose mean μ and variance σ^2 are given by (7) and (8), respectively. Thus, the expected number of backorders at any point in time is

$$(9) \qquad \sigma\phi\left(\frac{\mu}{\sigma}\right) - \mu\Phi\left(-\frac{\mu}{\sigma}\right).$$

which can easily be obtained by evaluating

$$\int_{-\infty}^{0} xh(x)\,dx.$$

Since

$$E(\text{on-hand inventory}) = E(\text{inventory position}) + E(\text{backorders})$$
$$- E(\text{number in repair}) - E(\text{number on order}),$$

the expected on-hand inventory is equal to

$$(10) \qquad r + 1 + \frac{Q-1}{2} + \frac{\gamma}{\lambda-\gamma} + \sigma\phi\left(\frac{\mu}{\sigma}\right) - \mu\Phi\left(-\frac{\mu}{\sigma}\right) - E(R(t)) - (\lambda-\gamma)\tau.$$

Note that the last term, the expected amount on order at any point in time, is equal to the rate at which demands are ultimately met by outside procurement, $\lambda - \gamma$, times the constant procurement lead time, τ.

In what follows, it will be easier to think of μ and σ^2 as functions of r and Q. Specifically, let

$$(11) \qquad \mu = r + \frac{Q}{2} + c,$$

and

$$(12) \qquad \sigma^2 = \frac{Q^2}{12} + d,$$

where

$$(13) \qquad c = \frac{\gamma}{\lambda-\gamma} + \frac{1}{2} - E(R(t)) - (\lambda-\gamma)\tau$$

and

$$(14) \qquad d = \frac{\lambda\gamma}{(\lambda-\gamma)^2} - \frac{1}{12} + \text{Var}(R(t)) + (\lambda+\gamma)\tau,$$

where we have used the approximation that $\text{Var}(Z(t-\tau, t)) = \gamma\tau$. This approximation is exact for $M|M|s$ and $M|G|\infty$ queueing systems. Note also that the constants c and d are independent of r and Q, and that the restriction that Q be greater than or equal to one guarantees that σ^2 is positive.

Finally, the rate at which demands are met by outside procurement, $\lambda - \gamma$, divided by Q, the procurement lot size, gives the expected number of procurement orders placed per year.

Combining our previous results, we see that the optimization problem for finding the optimal Q and r is

$$(15) \qquad \underset{Q \geq 1,\, r \geq 0}{\text{minimize}}\; K = \frac{(\lambda-\gamma)A}{Q} + (\hat{\pi} + h)\left[\sigma\phi\left(\frac{\mu}{\sigma}\right) - \mu\Phi\left(-\frac{\mu}{\sigma}\right)\right]$$
$$+ h\left[r + \frac{1}{2}Q + c\right],$$

where c is given by (13). This formulation of the problem came as a result of a number of key assumptions and approximations, which we now summarize:

(a) The demand and return processes are independent Poisson processes.

(b) The return rate is less than the demand rate.

(c) A continuous review (Q,r) policy is followed.

(d) The procurement lead time in constant.

(e) All demand not immediately satisfied is backordered.

(f) The distribution of net inventory is approximated by a normal distribution whose mean and variance are given by (7) and (8), respectively.

The objective function K is not convex in Q, but is convex in r. This is easily proven by noting that the backorder function $\sigma\phi\left(\frac{\mu}{\sigma}\right) - \mu\Phi\left(-\frac{\mu}{\sigma}\right)$ is convex in μ, $r = \mu - Q/2 \quad c$, and r is not related to σ. Thus, the optimal value of r satisfies $\frac{\partial K}{\partial r} = 0$, that is,

$$\Phi\left(-\frac{\mu}{\sigma}\right) = \frac{h}{\hat{\pi} + h}. \tag{16}$$

Thus, for a fixed value of Q, the variance of the normal distribution representing net inventory is fixed. Only the mean, or "location" of the curve, is decided by choosing a value of r. Therefore, Equation (16) indicates that once the variance is fixed, the "location" of the normal curve should be chosen so that the cumulative area to the left of the y-axis is $\frac{h}{\hat{\pi} + h}$, as illustrated in Figure 3.

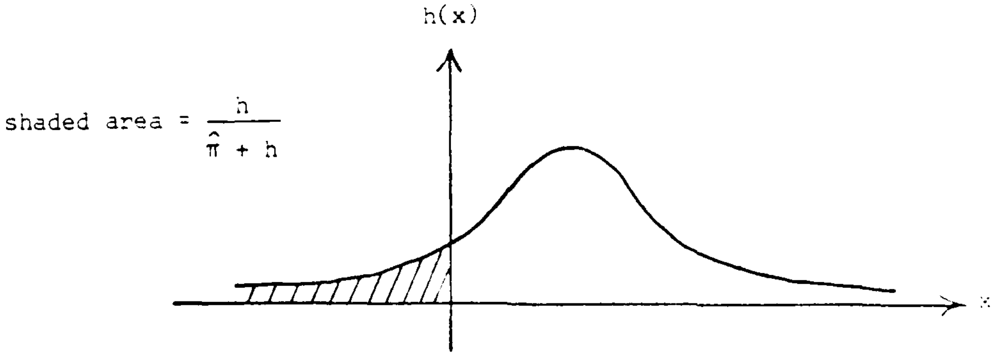


FIGURE 3. Location of the normal curve

In most real situations, the backorder cost $\hat{\pi}$ is large compared to the holding cost h. This makes the fraction $\frac{h}{\hat{\pi} + h}$ small. Recall that this fraction is the area to the left of the y-axis under the normal curve. The expected number of backorders is calculated using Equation (9), and the expected on-hand inventory is calculated in Equation (10) also using (9). Thus, as we stated earlier, accuracy of the normal approximation is required only in the tail of the distribution, since $\frac{h}{\hat{\pi} + h}$ is usually small.

Returning to (16) and rewriting it in terms of r^* and Q^*, the optimal values of r and Q, respectively, we have

$$\Phi\left(\frac{r^* + \dfrac{Q^*}{2} + c}{\sqrt{\dfrac{(Q^*)^2}{12} + d}}\right) = \frac{\hat{\pi}}{\hat{\pi} + h}$$

or

$$(17) \qquad r^* = \sqrt{\frac{(Q^*)^2}{12} + d} \cdot \Phi^{-1}\left(\frac{\hat{\pi}}{\hat{\pi} + h}\right) - \frac{1}{2}Q^* - c.$$

For a fixed value of Q, the optimal value of r is given by Equation (17).

To find the optimal value of Q, one can rewrite Equation (15) in terms of r and Q. Using Equation (17) to write the objective function solely as a function of Q, (15) simplifies to

$$K = \frac{(\lambda - \gamma)A}{q} + (\hat{\pi} + h)\sqrt{\frac{Q^2}{12} + d} \cdot \phi\left(\Phi^{-1}\left(\frac{\hat{\pi}}{\hat{\pi} + h}\right)\right).$$

This can be seen to be a convex function of Q. While the original objective function, K, is not convex everywhere in both Q and r, upon deriving an optimality condition (17), K is convex in both Q and r over the region of interest. Setting $\dfrac{dK}{dQ} = 0$, we find that Q^* is the value of Q that satisfies

$$(19) \qquad \frac{Q^3}{\sqrt{\dfrac{Q^2}{12} + d}} = \frac{12(\lambda - \gamma)A}{\alpha},$$

where

$$\alpha = (\hat{\pi} + h)\phi\left(\Phi^{-1}\left(\frac{\hat{\pi}}{\hat{\pi} + h}\right)\right).$$

If $Q^* < 1$, then set $Q^* = 1$.

Note that in realistic situations $d > 0$ (see Equation (14)), so the left side of (19) should increase with Q. A search method, such as either the Fibonacci or binary search technique, can be used to find Q^* in this case. Note, also, the similarity to the usual lot size formula. Ignoring some of the constants, (19) is roughly of the form

$$Q = \sqrt{\frac{(\lambda - \gamma)A}{h} \cdot \text{constant}}.$$

Also, observe that (19) is independent of r. Thus, once Q^* is found, r^* is found using (17).

3. THE MULTI-ECHELON CASE

In this section we study a two echelon system, which corresponds to the real system examined by the authors. The upper echelon consists of a warehouse having both repair and storage facilities that support the N lower echelon retailers. The retailers only have storage facilities.

All primary customer demands and returns are assumed to occur only at the retailers. We again assume that all customer demands not immediately satisfied are backordered, and that the demand and return processes are mutually independent Poisson processes. We also assume that lateral resupply is not allowed between retailers.

Let

λ_j = the customer demand rate at retailer $j (j = 1, \ldots, N)$,

γ_j = the customer return rate at retailer $j (j = 1, \ldots, N)$,

T_1 = the constant transportation time between the warehouse and a retailer, and

T_2 = the constant procurement lead time for the warehouse from an outside source.

The assumptions that transportation times are identical between the warehouse and any of the retailers, and that customer demands and returns occur only at the retailers are made for notational simplicity only. It will be apparent that relaxing these assumptions poses no additional problems.

Recall that repair facilities exist only at the upper echelon. Consequently, we assume that when a customer returns a repairable unit to a retailer it is immediately sent to the warehouse from the retailer and need not go back to that same retailer after it is repaired. We also assume that the repair process at the warehouse operates as a first-come, first-served queueing system.

Since transportation times are assumed to be constant, returns of repairable units to the warehouse occur as a Poisson process with rate $\gamma_0 = \sum_{j=1}^{N} \gamma_j$. Therefore, it is equivalent, and more convenient, to think of returns occurring only to the warehouse, and as a Poisson process with rate γ_0.

We assume that retailer j uses an $(S_j - 1, S_j)$ continuous review ordering policy, i.e., a constant inventory position (net inventory plus on order) of S_j is maintained. This implies that retailer j immediately orders one unit from the warehouse every time a customer demand occurs at the retailer. Since each order placed at a retailer also results in a demand being placed upon the warehouse, demands on the warehouse occur as a Poisson process with rate $\lambda_0 = \sum^{N} \lambda_j$.

[Note the importance of the assumption of following an $(S_j - 1, S_j)$ policy at retailer j. If the retailers followed (Q,r) ordering policies, then the time between the placing of orders upon the warehouse would not necessarily be exponential, nor would the orders necessarily be for individual units. Thus, the demand process at the warehouse would no longer be a simple Poisson process.]

We assume that $\gamma_0 < \lambda_0$ so that an occasional outside procurement is necessary. The warehouse is assumed to follow a (Q_0,r_0) policy, i.e., when its inventory position (net inventory plus on order plus in repair) falls below $r_0 + 1$, an order for Q_0 units is placed upon an outside procurement source.

Warehouse procurement orders are assumed to arrive at the warehouse T_W time units after the order is placed. However, an order placed by a retailer upon the warehouse does not necessarily arrive at the retailer T_j time units after it is placed. In addition to the transportation time, there may be a delay due to the warehouse being out of serviceable stock. All demands made upon the warehouse that are not immediately satisfied are backordered.

A schematic representation of this system is given by Figure 4

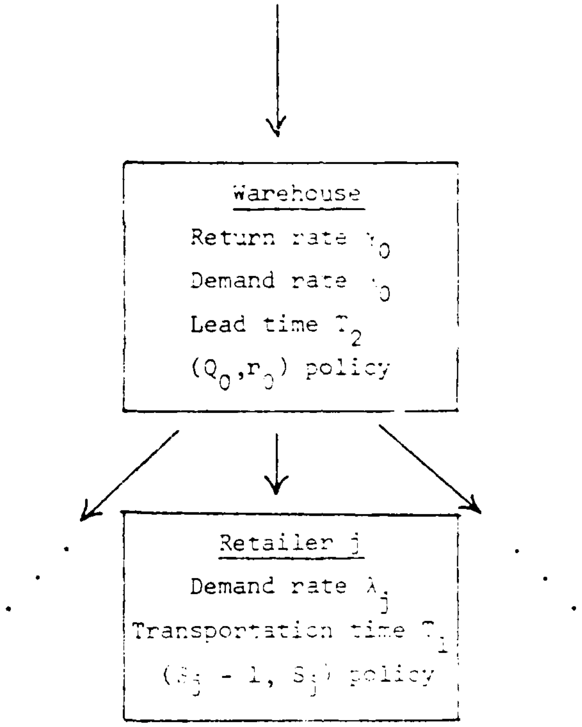


FIGURE 4. Schematic representation of the multi-echelon system

Finally, let the system cost parameters be as follows:

h_0 = the holding cost at the warehouse (\$/unit-year),

h_j = the holding cost at retailer j (\$/unit-year) ($j = 1, \ldots, N$),

$\hat{\pi}_j$ = the backorder cost at retailer j (\$/unit-year) ($j = 1, \ldots, N$),

and A = the fixed warehouse procurement order cost (\$/order).

Given values of $h_j (j = 0, \ldots, N)$, $\hat{\pi}_j (j = 1, \ldots, N)$, and A, all assumed to be nonnegative, the problem is to determine values for Q_0, r_0, and S_j ($j = 1, \ldots, N$) that will minimize the expected annual sum of the retailer holding and backorder costs, and the warehouse ordering and holding costs. Thus, the optimization problem we want to solve is

$$(20) \qquad \min_{Q_0, r_0, S} \left\{ \sum_{j=1}^{N} (h_j \cdot E\{\text{On-hand Inventory at Retailer } j\} + \hat{\pi}_j \cdot E\{\text{Backorders Outstanding at a Random Point in Time at Retailer } j\}) + A \cdot \frac{\lambda_0 - \gamma_0}{Q_0} + h_0 \cdot E\{\text{On-hand Inventory at the Warehouse}\} \right\}$$

subject to $Q_0 \geqslant 1$, $r_0 \geqslant 0$ and $S_j = 0, 1, \ldots,$ for $j = 1, \ldots, N$.

The expected on-hand inventory at the warehouse can be found using Equation (10); however, the expected on-hand inventory and backorders at retailer j cannot be determined as easily. We will subsequently show how these expectations can be calculated.

Note that we have not explicitly stated a value for $\hat{\pi}_0$, the warehouse backorder cost, and that this cost is not included in the objective function that is to be minimized. Given the interactions between the two echelons of our inventory system, the cost of a backorder at the warehouse is not an explicit one but rather an imputed one. It is measured by the effect of a backorder at the upper echelon on the expected performance at the lower echelon.

The optimal stock level at retailer j, S_j^*, is a function of the procurement resupply time, that is, the expected time from the placement to receipt of an order by a retailer. This procurement resupply time is then the transportation time, T_1, plus the expected delay due to the warehouse being out of serviceable stock. Clearly, costs at the retail echelon can be lowered by reducing the expected resupply time. *This can only be accomplished* by decreasing the expected warehouse backorders at a random point in time, which is achieved by increasing Q_0 or r_0 (or both). This, in turn, raises holding costs at the warehouse. Thus, a tradeoff exists between holding costs at the upper echelon and holding and backorder costs at the lower echelon. We will present an iterative algorithm based on this tradeoff which alternates between finding stock levels for the upper and lower echelons. The basis for this algorithm, presented in Section 3.1, is founded on the results developed in Section 2.

3.1 Analysis

Suppose the imputed cost of a warehouse backorder is known to be $\hat{\pi}_0$. Then we can use (17) and (19) to find optimal values for r_0 and Q_0. These determine a "performance level" B, where

B = the expected backorders at the warehouse at a random point in time

$$= \sigma\phi\left(\frac{\mu}{\sigma}\right) - \mu\Phi\left(-\frac{\mu}{\sigma}\right),$$

and where μ and σ^2 are the mean and variance, respectively, of the normal approximation to the stationary distribution of net inventory at the warehouse.

Then the expected resupply time for a retailer is

$$(21) \qquad T = T_1 + B/\lambda_0.$$

since the expected delay time per demand is the expected number of backorders at a random point in time divided by the demand rate. This is a direct application of Little's Formula $L = \lambda W$. Then, using Palm's Theorem [1] as an approximation, we assume the number of units in resupply at retailer j to be Poisson distributed with mean $\lambda_j T$.

Note: Palm's Theorem requires the independence of resupply times, making this system analogous to an $M/G/\infty$ queue. Resupply times in our system are not independent; consider, for example, a demand by a retailer which cannot be immediately filled by the warehouse. Then it is more likely that the next demand placed by a retailer upon the warehouse also experiences a delay than if the preceding order had been immediately satisfied. This approximation of the distribution of the number of units in resupply at retailer $j (j = 1, \ldots, N)$ was tested for the special case in which the repair facility at the warehouse behaves as an $M/D/\infty$ queueing system. The exact distribution of $R_j(t)$, the number of units in resupply at retailer j, was obtained from comparison with the Poisson approximation. Our analysis indicates that the Poisson approximation improves as the expected warehouse backorders, or the probability of delay at the warehouse, decreases. In particular, the Poisson approximation was found to be good as long as the expected value of net inventory at the warehouse at a random point in time is greater than zero. (In the test cases in which this condition was met the maximum absolute difference between $R_j(t)$ and its Poisson approximation was less than 5%.) This will, of course, be the case for a reasonably large ratio of backorder to holding costs.

Once we know the value of T and have the form (approximately) of the distribution of the number of units on order by retailer j, we can solve N independent subproblems to obtain the optimal value for S_j. The subproblem at retailer j consists of finding the optimal stock level S_j^*, assuming a constant procurement resupply time of T, where T is given by (21). This is accomplished using Lemma 1.

LEMMA 1: Suppose the procurement lead time is a constant T and demand is Poisson distributed with rate λ_j. Then the optimal value S_j^* for an $(S_j - 1, S_j)$ policy is the largest integer S_j such that

$$(22) \qquad \underline{P}(S_j;\lambda_j T) > \frac{h_j}{\hat{\pi}_j + h_j},$$

where $$\underline{P}(x,\mu) = \sum_{r=x}^{\infty} p(r;\mu)$$

and $$p(r;\mu) = e^{-\mu}\frac{\mu^r}{r!}.$$

The proof of Lemma 1 can be found on page 204 of Reference 5.

Let $K_j(S_j, T)$ be the expected annual holding and backorder costs at retailer j when the inventory position is S_j and the procurement lead time is a constant T. As can be shown (see Reference 5)

$$(23) \qquad K_j(S_j,T) = (\hat{\pi}_j + h_j)[\lambda_j T\underline{P}(S_j - 1;\lambda_j T) - S_j\underline{P}(S_j;\lambda_j T)] + h_j[S_j - \lambda_j T\}.$$

For a fixed value of T (and therefore of B) we define the minimum total expected costs at the lower echelon, $K^L(B)$, as

$$(24) \qquad K^L(B) = \sum_{j=1}^{N} K_j(S_j^*, T).$$

where $K_j(\cdot, \cdot)$ is given by (23), T is given by (21), and S_j^* satisfies (22).

Note that when $B = b, \left.\dfrac{dK^l(B)}{dB}\right|_{B=b}$ is an estimate of $\hat{\pi}_0$, since it measures the marginal effect of a warehouse backorder on the expected total lower echelon cost. It is easy to show that

$$(25) \qquad \frac{dK^l(B)}{dB} = \frac{1}{\lambda_0} \sum_{j=1}^{N} [(\hat{\pi}_j + h_j)\lambda_j \underline{P}(S_j^*; \lambda_j T) - h_j \lambda_j].$$

Next, let $K^w(B)$ represent the minimum expected warehouse ordering and holding cost given that B, the expected number of warehouse backorders outstanding at a random point in time, is fixed. In particular, we define

$$K^w(B) = \min_{\substack{Q_0 \geq 1 \\ r_0 \geq 0}} \left\{ \frac{\lambda_0 - \gamma_0}{Q_0} \cdot A + h_0 \cdot \left[r_0 + \frac{Q_0}{2} + B + c \right] \right\}$$

subject to $B = \lambda_0(T - T_1)$,

where the constant c is given by Equation (13).

We conclude this section with the statement of two additional lemmas

LEMMA 2: $K^w(B)$ is convex decreasing in B.

LEMMA 3: Let T be a constant resupply time. If the optimal stock levels S_j ($j = 1, \ldots, N$) are continuous rather than integer valued, then $K^l(B)$ is a concave increasing function of B, where $B = \lambda_0(T - T_1)$. These lemmas can be proved by applying the chain rule to take derivatives. The details can be found in Reference 9.

3.2 Restatement of Problem 20

Problem (20) can be restated based on the interrelationship between the warehouse and the retailers developed in Section 3.1. As we have demonstrated, the two echelons are linked through the value of B. Then an alternative way of writing problem (20) is

$$(26) \qquad \min_{B \geq 0} K^l(B) + K^w(B)$$

where $\qquad B = \lambda_0(T - T_1)$.

Figure 5 represents a typical graphing of $K^l(B)$ and $K^w(B)$ as functions of B. We observed in all test cases that, under the conditions of Lemma 3, $K^w(B) + K^l(B)$ was a convex function of B. Thus, the minimum cost will occur where

$$(27) \qquad \frac{dK^l(B)}{dB} = -\frac{dK^w(B)}{dB}.$$

The algorithm presented in the next section takes advantage of the fact that problem (20) can be restated as problem (26) and that the optimal solution must satisfy (27).

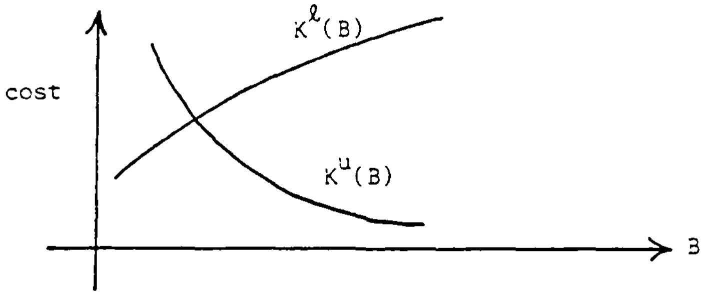


FIGURE 5. Minimum upper and lower echelon cost functions vs. B.

3.3 An Algorithm

The following algorithm can be used to solve problem (26):

STEP 0: Let $\hat{\pi}_0 = \max\limits_{j=1,\ldots,N} (\pi_j)$.

STEP 1: Given $\hat{\pi}_0$, solve for Q_0, r_0 using Equations (17) and (19, and determine the corresponding value of B, say b.

STEP 2: Let $T = T_1 + b/\lambda_0$; find the S_j^* using Equation (22).

STEP 3: Using these S_j^*, find $\dfrac{dK^l(B)}{dB}$ evaluated at $B = b$, using Equation (25); let $\hat{\pi}_0$ assume this value, and return to Step 1 unless the stock levels and costs have converged sufficiently.

The first few steps of the above algorithm are illustrated in Figure 6. The algorithm begins by setting $\hat{\pi}_0 = \max\limits_{j=1,\ldots,N} (\hat{\pi}_j)$. This is an upper bound on the optimal value of $\hat{\pi}_0$, since this value implies that a backorder at the warehouse always results in a backorder at the retailer with the largest backorder cost. Then Q_0 and r_0 are found using this upper bound on $\hat{\pi}_0$. This determines a value of B (say $B = b_1$) (and therefore of T), which is a lower bound on the optimal value of B (and therefore of T). These computations yield point ① on the upper echelon cost curve in Figure 6.

Using this lower bound on the optimal value of T, we find a lower bound estimate of S_j^* ($j = 1, \ldots, N$), which determines a value $K^l(b_1)$, and point ② in Figure 6. Next we set $\hat{\pi}_0 = \left.\dfrac{dK^l(B)}{dB}\right|_{B-b_1}$. Since $K^l(B)$ is concave in B, and since we have a lower bound estimate of the optimal B, the new estimate of $\hat{\pi}_0$ is an upper bound on the optimal value of $\hat{\pi}_0$; but it is smaller than the previous estimate. Using this new estimate of $\hat{\pi}_0$, B will increase to a value, say b_2, as a result of resolving for r_0 and Q_0 using (17) and (19). These calculations produce point ③ in Figure 6. The procedure continues by letting $T = T_1 + \dfrac{b_2}{\lambda_0}$ and finding $K^l(b_2)$, which leads to point ④. The algorithm continues in this manner until convergence occurs. Discussion of convergence and other aspects of the algorithm can be found in Reference 9.

The algorithm was tested on 50 problems. In general, the values of Q_0^*, r_0^* and S_j^* ($j = 1, \ldots, N$) were found after only three iterations of the algorithm. This occurred in 48 of the 50 test cases. The curve $K^l(B)$ is very flat compared to $K^u(B)$, so that convergence to the

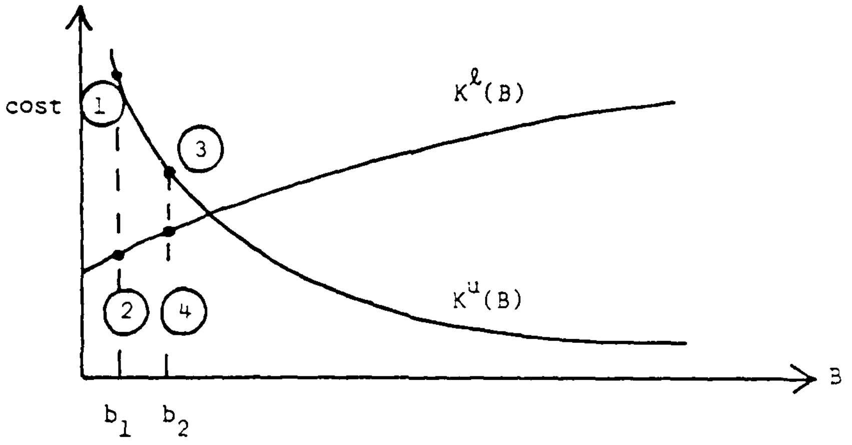


FIGURE 6. First steps of the algorithm

correct value of $\tilde{\pi}_0$ occurs quickly. As we noted earlier, $K^l(B) + K^u(B)$ was convex for all of the 50 test problems. The reason this occurred was because $K^l(B)$, although concave, is almost linear.

4. SUMMARY AND CONCLUDING COMMENTS

We have developed simple methods for obtaining parameter values for a procurement policy for certain inventory systems with returns. The key was the use of a normal approximation to the steady-state distribution of net inventory. This led to the development of cost models which were easily solved.

In the single location model, we assumed the procurement policy to be a stationary (Q,r) policy. This policy is not the optimal one. In Reference 9 it is shown that, for the special cases of $M/M/1$ and $M/G/\infty$ queueing repair systems for which the transient distributions of the repair system's output are easily developed, one can lower total expected costs by redefining inventory position and allowing variable reorder points as follows. Inventory position is redefined to be net inventory plus the number of units on order. The analysis proceeds exactly as described in Section 2 (with some of the constants redefined). This results in a reduction in σ^2, the variance of net inventory, since the variance of the number of units in repair is no longer included in σ^2. The reorder point, expressed in terms of inventory position, is then a function of the number of units in repair, rather than a constant. Reductions in total expected costs can be achieved by using a state dependent reorder point when the variance of the number of units in repair is very large. A 10% reduction in total expected cost was achieved using the variable reorder point policy in an $M/M/1$ repair system with traffic intensity $\rho = 499/500$. This is an extreme case, however. The average annual cost of using the stationary (Q,r) policy was within 1% of the average annual cost obtained using the nonstationary one in almost all test cases. Since this is the case, and since a stationary (Q,r) policy is easy to use, the stationary (Q,r) policy is an attractive policy to implement.

Next, we showed how the single location solution method can be incorporated into an iterative algorithm for setting stock levels in the single item, multi-echelon inventory problem with returns. The algorithm can also be extended to find stock levels in an M-echelon inventory system with returns. The only requirement would be that an $(S-1, S)$ procurement policy must be followed at each of the lower $M-1$ echelon locations.

REFERENCES

[1] Feeney, G.J. and C.C. Sherbrooke, "The $(S-1, S)$ Inventory Policy Under Compound Poisson Demand," Management Science, *12*, 391-411 (1966).

[2] Gajdalo, S., "Heuristics for Computing Variable Safety Levels/Economic Order Quantities for Repairable Items," AMC Inventory Research Office, Institute of Logistics Research, U.S. Army Logistics Management Center, Fort Lee, VA (1973).

[3] Gross, D. and C.M. Harris, *Fundamentals of Queueing Theory*, (John Wiley and Sons, New York, 1974).

[4] Gross, D., H.D. Kahn and J.D. Marsh, "Queueing Models for Spares Provisioning," Naval Research Logistics Quarterly, *24*, 521-536 (1977).

[5] Hadley, G. and T.M. Whitin, *Analysis of Inventory Systems*, (Prentice-Hall, New Jersey, 1963).

[6] Heyman, D.P, "Optimal Disposal Policies for a Single-Item Inventory System with Returns," Naval Research Logistics Quarterly, *24*, 385-405 (1977).

[7] Heyman, D.P , "Return Policies for an Inventory System with Positive and Negative Demands," Naval Research Logistics Quarterly, *25*, 581-596 (1978).

[8] Hoadley, B. and D.P. Heyman, "A Two-Echelon Inventory Model with Purchases, Disposition, Shipments, Returns, and Transshipments," Naval Research Logistics Quarterly, *24*, 1-19 (1977).

[9] Isaac, M.H., "An Analysis of Inventory Systems with Returns," unpublished Ph.D. dissertation, School of Operations Research and Industrial Engineering, Cornell University (1979).

[10] Miller, B.L., "Dispatching from Depot Repair in a Recoverable Item Inventory System: On the Optimality of a Heuristic Rule," Management Science, *21*, 316-325 (1974).

[11] Muckstadt, J.A., "A Model for a Multi-Item, Multi-Echelon, Multi-Indenture Inventory System," Management Science, *20*, 472-481 (1973).

[12] Porteus, E.L. and Z. Lansdowne, "Optimal Design of a Multi-Item, Multi-Location, Multi-Repair Type Repair and Supply System," Naval Research Logistics Quarterly, *21*, 213-237 (1974).

[13] Ross, S.M., *Introduction to Probability Models*, (Academic Press, New York, 1972).

[14] Schrady, D.A., "A Deterministic Inventory Model for Repairable Items," Naval Research Logistics Quarterly, *14*, 391-398 (1967).

[15] Sherbrooke, C.C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control," Operations Research, *16*, 122-141 (1968).

[16] Simpson, V.P., "Optimum Solution Structure for a Repairable Inventory Problem," Operations Research, *26*, 270-281 (1978).

ANALYTIC APPROXIMATIONS FOR (s,S) INVENTORY POLICY OPERATING CHARACTERISTICS*

Richard Ehrhardt

Curriculum in Operations Research and Systems Analysis
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

ABSTRACT

The operating characteristics of (s,S) inventory systems are often difficult to compute, making systems analysis a tedious and often expensive undertaking. Approximate expressions for operating characteristics are presented with a view towards simplified analysis of systems behavior.

The operating characteristics under consideration are the expected values of: total cost per period, period-end inventory, period-end stockout quantity, replenishment cost per period, and backlog frequency. The approximations are obtained by a two step procedure. First, exact expressions for the operating characteristics are approximated by simplified functions. Then the approximations are used to design regression models which are fitted to the operating chracteristics of a large number of inventory items with diverse parameter settings. Accuracy to within a few percent of actual values is typical for most of the approximations.

1. INTRODUCTION

There are many situations in which an inventory system's designer can use estimates of operating characteristics of the system. For example, management may desire forecasts of inventory on hand, or system operating costs. Our goal in this paper is to develop simple approximations that designers can use to estimate the following operating characteristics of a periodic-review inventory system: average holding cost per period, average backlog cost per period, frequency of periods without backlogs, average replenishment cost per period, and average total cost per period. These characteristics are defined mathematically in Section 2.

We consider a periodic-review, single-item inventory system where backlogging is permitted and there is a fixed lead time between placement and delivery of an order. Demands during review periods are represented by independent, identically distributed random variables having mean μ and variance σ^2. Replenishment costs are composed of a setup cost K and a unit cost c. There is a fixed lead time L between the placement and delivery of each replenishment order. At the end of each review period, a cost h or p is incurred per unit on hand or backlogged, respectively. The criterion of optimality is minimization of the expected undiscounted cost per period over an infinite horizon.

Under these assumptions it has been shown that there exists an optimal policy of the (s,S) form (Iglehart [3]). That is, a replenishment order is not placed unless the inventory position (on-hand plus on-order minus backorders), x, is less than or equal to s, at which time

*This research was supported by contracts with the Office of Naval Research and the U. S. Army Research Office

an order of size $S - x$ is placed. Computational methods have been developed (Veinott and Wagner [6]) for calculating optimal policies and their operating characteristics. Unfortunately, the computational effort required is prohibitive for practical implementation. Furthermore, exact computation requires the complete specification of the demand distribution, a level of detailed information that is unlikely to be available in practice.

In this paper we develop approximations for operating characteristics in a two step procedure. We start with exact analytic expressions for the operating characteristics and approximate the exact expressions with simplified functions. Then we generalize the functions and fit their parameters to the observed characteristics of 576 items using least-squares regression. The resulting approximations are accurate and require for demand information only the mean and variance. In Section 2 we derive the simplified functions from exact expressions for the operating characteristics, and in Section 3 we present the results of the regression analyses. Finally, in Sections 4 and 5 we analyze the accuracy of the approximations and draw conclusions.

2. ANALYTIC APPROXIMATIONS

Consider the model of Section 1 and assume that demand follows a probability density $\phi(\cdot)$ and cumulative distribution $\Phi(\cdot)$. Let $\phi^{*n}(\cdot)$ and $\Phi^{*n}(\cdot)$ be the n-fold convolutions of these functions. We consider the following operating characteristics of fixed, infinite-horizon (s,S) policies:

(1) $H \equiv$ average holding cost per period,

$B \equiv$ average backlog cost per period,

$P \equiv$ backlog protection, i.e., frequency of periods without backlogs,

$R \equiv$ average replenishment cost per period, and

$T \equiv$ average total cost per period.

Let

$$m(\cdot) \equiv \sum_{n=1}^{\infty} \phi^{*n}(\cdot),$$

and

$$M(\cdot) \equiv \sum_{n=1}^{\infty} \Phi^{*n}(\cdot).$$

The functions $m(\cdot)$ and $M(\cdot)$ are renewal functions which govern the frequency of replenishments, and, therefore, the evolution of the inventory positions. We have, as in Roberts [4], the exact relationships

$$(2)\qquad \begin{aligned} H &= h[1 + M(D)]^{-1}\left\{\int_0^D \int_0^{S-y} (S - y - x)\phi^{*(L+1)}(x)m(y)\,dx\,dy \right.\\ &\qquad \left. + \int_0^S (S - x)\phi^{*(L+1)}(x)\,dx\right\} \\ B &= p[H/h + (L + 1)\mu - S] + p[1 + M(D)]^{-1}\int_0^D y\,m(y)\,dy \\ P &= [1 + M(D)]^{-1}\left\{\int_0^D \Phi^{*(L+1)}(S - y)m(y)\,dy + \Phi^{*(L+1)}(S)\right\} \\ R &= K[1 + M(D)]^{-1} \\ T &= H + B + R. \end{aligned}$$

where

$$D \equiv S - s.$$

Notice that a constant term $c\mu$ has been omitted from the expression for replenishment cost P since it does not affect the choice of an optimal policy. It is difficult to obtain any insights from (2) regarding the sensitivity of the operating characteristics to values of model parameters. Indeed, it is exceedingly complicated just to calculate values of the characteristics for a given set of parameter values. We proceed to simplify the form of expression (2) by introducing approximations for the functions $m(\cdot)$, $M(\cdot)$, and $\phi^{*(L+1)}(\cdot)$.

Replenishment frequency in (2) is given by $[1 + M(D)]^{-1}$. To approximate $M(\cdot)$, we use the following result of Smith [5]:

$$M(x) = x/\mu + \sigma^2/(2\mu^2) - 1/2 + o(1), \; x \to \infty.$$

This yields the approximate value for replenishment frequency

$$(3) \qquad [1 + M(D)]^{-1} \doteq \mu/[D + (\mu + \sigma^2/\mu)] \equiv \rho.$$

To obtain approximations for the other characteristics in (2), we first need to find a simple expression for the function $m(\cdot)$. We identify the quantity $(S - y)$ in (2) as the inventory position (after ordering), with stationary distribution function $F(\cdot)$ given by

$$F(S - y) = \begin{cases} M(y)/[1 + M(D)] \, , & s \leqslant S - y < S \\ 1 & , \; S - y = S. \end{cases}$$

The probability density $f(\cdot)$ of the inventory position (after ordering) on the interval $[s,S)$ is

$$f(S - y) = m(y)/[1 + M(D)].$$

We approximate $f(\cdot)$ by a constant c on the interval $[S,s)$. There are two reasons why this should be a reasonable approximation. Firstly, the result of Smith [5] shows that $m(y)$ is asymptotically constant as y grows large. Secondly, we know that $f(\cdot)$ is exactly constant for the special case of an exponential demand distribution.

We find a value for c by normalizing the approximated distribution. Starting with an exact expression, we have

$$(5) \qquad \int_s^S f(S - y)dy = M(D)/[1 + M(D)].$$

Then we substitute c for $f(\cdot)$ on the left side of (5) and use (3) on the right side of (5), yielding

$$(6) \qquad c = (1 - \rho)/D.$$

We use (3) and (6) in (2) to get

$$(7) \qquad H \doteq h\left\{[(1 - \rho)/D] \int_0^D \int_0^{S-y} (S - y - x)\phi^{*(L+1)}(x)dxdy \right.$$
$$\left. + \rho \int_0^S (S - x)\phi^{*(L+1)}(x)dx\right\}$$

$$B \doteq \rho[H/h + (L + 1)\mu - S + (1 - \rho)D/2]$$

$$P \doteq \rho\,\Phi^{*(L+1)}(S) + [(1 - \rho)/D] \int_0^D \Phi^{*(L+1)}(S - y)dy$$

$$R \doteq \rho K.$$

The expressions for H, B, and P in (7) still require the specification of the demand distribution. We obtain a further simplification by approximating the demand distribution with a gamma distribution. As we show below, this approximation leads to expressions for H, B, and P that require for demand information only the mean μ and variance σ^2. The class of gamma distributions provides good fits for a wide variety of unimodal or nonincreasing densities on the positive real line and should be a reasonable approximation in our application. For inventory items that have significantly non-gamma demand distributions, an analyst could produce a new set of approximations by making the appropriate substitution in (7) and proceeding in the manner described below.

Let $g(\cdot|\alpha,\beta)$ be a gamma density function with parameters α and β. Then we have

(8) $$\phi^{*(L+1)}(x) \doteq g(x|\alpha,\beta) = \begin{cases} x^{\alpha-1}\exp(-x/\beta)/[\Gamma(\alpha)\beta^{\alpha}] , & x \geqslant 0 \\ 0 , & x < 0 \end{cases}$$

$$\Phi^{*(L+1)}(x) \doteq G(x|\alpha,\beta) \equiv \int_0^x g(y|\alpha,\beta)dy,$$

where

$$\alpha \equiv (L+1)\mu^2/\sigma^2$$

$$\beta \equiv \sigma^2/\mu$$

We define the notation

$$[f(x)]_a^b \equiv f(b) - f(a),$$

and use (8) in (7) to yield

(9) $$H \doteq ph\{SG(S|\alpha,\beta) - \alpha\beta G(S|\alpha+1,\beta)\}$$
$$+ [h(1-p)/2D]\,[x^2G(x|\alpha,\beta) - 2\alpha\beta xG(x|\alpha+1,\beta)$$
$$+ (\alpha+1)\alpha\beta^2\,G(x|\alpha+2,\beta)\,]_s^S$$
$$B \doteq p\{H/h + (L+1)\mu - S + (1-p)D/2\}$$
$$P \doteq pG(S|\alpha,\beta) + [(1-p)/D][xG(x|\alpha,\beta) - \alpha\beta G(x|\alpha+1,\beta)\,]_s^S$$
$$R \doteq pK.$$

Observe that the approximations (9) depend on the values of s,S, the economic parameters, and the mean and variance of demand. The function G must be calculated by a numerical procedure. We use a series expansion for $G(x|\alpha,\beta)$ when x is less than the minimum of 1 and $\alpha\beta$, and a continued-fraction expansion otherwise. The procedure is part of a package of computer programs entitled "The IMSL Library" which is marketed by the International Mathematical and Statistical Libraries, Inc., Houston, Texas.

Despite the effort required to compute G, the expressions in (9) are an enormous simplification over (2). In Section 5 we mention the possibility of using a normal distribution function in lieu of the function G. Employing the normal distribution would facilitate manual computations of the approximations we derive below.

3. NUMERICAL ANALYSIS

In this section we use expressions (9) to develop regression models for the operating characteristics. We fit the parameters of the regression models to the observed characteristics of 576 items. The 576-item system is formed by using a full factorial combination of the parameters in Table 1. Discrete demand distributions are used in the analysis with means ranging from 2 to 16 and variances ranging from 2 to 144. Although the expressions in (9) are based on a continuous demand distribution, we will show that they can be used to approximate many of the characteristics of items with discrete distributions, which are more common in practice. Notice that all the items in Table 1 have a unit holding cost h of 1. Since the total cost function is linear in K, p, and h, we have used h as a normalizing parameter.

TABLE 1 — *System Parameters*

Factor	Levels	Number of Levels
Demand distribution	Poisson $(\sigma^2/\mu = 1)$ Negative Binomial $(\sigma^2/\mu = 3)$ Negative Binomial $(\sigma^2/\mu = 9)$	3
Mean demand (μ)	2, 4, 8, 16	4
Replenishment lead time (L)	0, 2, 4	3
Replenishment setup cost (K)	32, 64	2
Unit penalty cost (p)	4, 9, 24, 99	4
Unit holding cost (h)	1	1
Policy	Optimal policy, power approximation policy	2

The (s,S) policies used in the 576-item system are of two types: those with optimal values of s,S computed with the algorithm of Veinott and Wagner [6] and approximately optimal values of s,S computed by the power approximation algorithm of Ehrhardt [1]. For each item in the system we use the methods in [6] to compute exact values of the characteristics in (2) and use these as data for our regression analyses. The approximations we obtain are labelled with subscript "a" when they are used for all 576 items. Subscripts "a,p" or "a,o" are used to label expressions that apply only to power approximation or optimal policies, respectively.

We develop our regression adjusted approximations in the following subsections. In each subsection, we derive an approximation and assess its accuracy in the 576-item system. The measure of accuracy we use is the absolute value of the percentage difference between the exact and approximated values for individual items. We note here that the accuracy of the approximations appears to be even greater when the operating characteristics are aggregated over portions of the 576-item system. That is, there are essentially no systematic errors with respect to any of the model parameters. For a more detailed discussion of this point, see [2].

An Approximation for Replenishment Cost

We use (3) in (9) to obtain the expression for replenishment cost

$$R \doteq \mu K/[D + (\mu + \sigma^2/\mu)/2].$$

We manipulate the expression to form a linear regression model

$$(\mu K/R) = A_0 + A_1 D + A_2\mu + A_3(\sigma^2/\mu) + \epsilon.$$

where $A_0, \ldots, A_3$ are constants to be fit and ϵ is the error term. We use least-squares regression to fit the model to the system of 576 inventory policies in Table 1. That is, for each of these policies we use D, μ, and σ^2/μ as independent variables, and we use the exactly computed value of $\mu K/R$ as the dependent variable. The result is the following numerical approximation for R:

$$(10) \qquad R_a = K\mu/[D + (\mu + \sigma^2/\mu)/2 - .5121].$$

which has a coefficient of determination (fraction of variance explained) of 0.9999 for the quantity $\mu K/R$.

When used in the 576-item system, expression (10) is within 0.1% of actual values of R, on the average. The expression is accurate to within 2% for all but 2 items, with a maximum error of 2.5%.

An Approximation for Holding Cost

We can treat the unit holding cost as a redundant (normalizing) parameter in our model, and so we divide the holding cost expression in (9) by h yielding

$$\begin{aligned} H/h \doteq{} & p[SG(S|\alpha,\beta) - \alpha\beta G(S|\alpha+1,\beta)] \\ & + [(1-p)/2D]\,\{x^2G(x|\alpha,\beta) - 2\alpha\beta xG(x|\alpha+1,\beta) \\ & + (\alpha+1)\alpha\beta^2G(x|\alpha+2,\beta)\}\Big|_s^S. \end{aligned}$$

We take advantage of our improved estimate of replenishment frequency from (10) and replace p with

$$(11) \qquad r \equiv R_a/K = \mu/[D + (\mu + \sigma^2/\mu)/2 - .5121].$$

The result is a quantity that we denote as W, given by

$$\begin{aligned} (12) \qquad W \equiv{} & r[SG(S|\alpha,\beta) - \alpha\beta G(S|\alpha+1,\beta)] \\ & + [(1-r)/2D]\,\{x^2G(x|\alpha,\beta) - 2\alpha\beta xG(x|\alpha+1,\beta) \\ & + (\alpha+1)\alpha\beta^2 G(x|\alpha+2,\beta)\}\Big|_s^S. \end{aligned}$$

We calculated values of W in the 576-item system. We compared them with the actual values of H/h and found a systematic variation with respect to μ and σ^2/μ. This motivates the linear regression model

$$H/h = A_0 + A_1W + A_2\mu + A_3(\sigma^2/\mu) + \epsilon,$$

where $A_0, \ldots, A_3$ are constants to be fit and ϵ is the error term. We use least-squares regression to fit the model to the system of 576 items. The result is a coefficient of determination of 0.9999 for the approximation

$$(13) \qquad H_a = h(W - .1512\mu + .1684\sigma^2/\mu + .0689).$$

Expression (13) is within 0.7% of actual values of H, on the average, when used in the 576-item system. It is accurate to within 2% for 96% of the items, and within 4% for 99% of the items. Only one item produces an error in excess of 6%. This error is 9.2% for the item controlled with optimal values of (s,S), μ equal 2, σ^2 equal 18, p/h equal 4, K/h equal 32, and L equal 0. In general, the largest errors occur for high values of variance-to-mean ratio and low values of other parameters.

An Approximation for Backlog Protection

Backlog protection is defined as the frequency of periods without backlogs, that is, one minus the backlog frequency. Since it is a critical measure of service, it is of central interest to the inventory systems designer. Unfortunately, when (9) is used to construct regression models for backlog protection, very poor fits result. The highest coefficient of determination obtained using this approach is 0.68.

We revised the regression model to reflect a theoretical result. When demand is continuously distributed, an optimal policy yields $(p/h)/(1 + p/h)$ for backlog protection. When the demand distributions are discrete, $(p/h)/(1 + p/h)$ is a lower bound on P for optimal policies. It was observed in [1] that the power approximation and optimal policies differed in their backlog frequency performance. Therefore, we decided to fit the two policy rules separately.

We use the model

$$(1 + p/h)P = A_0 + A_1(p/h) + \epsilon,$$

which dramatically improves the fit. For optimal policies, the simple expression

$$P_{a,o} = (0.0857 + p/h)/(1 + p/h) \tag{14}$$

yields a coefficient of determination of 0.99999 for $(1 + p/h)P$. We have the same coefficient of determination for power approximation policies with

$$P_{a,p} = (0.0695 + p/h)/(1 + p/h). \tag{15}$$

When used in the 576-item system, expressions (14) and (15) are accurate to within 0.7% on the average. They are accurate to within 2% for 92% of the items and to within 4% for 98% of the items. All nine items with errors in excess of 4% have power approximation policies with a unit penalty cost of 4. The approximations are especially accurate for large unit penalty costs.

An Approximation for Total Cost

We obtain an expression for total cost by summing cost components H, B, and R, and using approximations (9) for B and R

$$\begin{aligned} T &= H + B + R \\ &\doteq (1 + p/h)\, H + p[(L + 1)\mu - S + (1 - \rho)D/2] + \rho K. \end{aligned}$$

We divide by h, replace ρ with r, as given by (11), and use approximation (12) for H to obtain

$$T/h = (1 + p/h)W + p/h\,[(L + 1)\mu - S + (1 - r)D/2] + rK/h. \tag{16}$$

As we discovered in obtaining a fit for holding cost, a group of related terms should be added to (16) to obtain a good fit to the system's data. The linear regression model we employed is

$$\begin{aligned} T/h = {} & A_0 + A_1 W + A_2(Wp/h) + A_3[(L + 1)\mu p/h] + A_4(Sp/h) \\ & + A_5(Dp/h) + A_6(rDp/h) + A_7(rK/h) + A_8(p/h) \\ & + A_9(rp/h) + A_{10}[(L + 1)\mu] + A_{11}S + A_{12}D + A_{13}(Dr) \\ & + A_{14}r + A_{15}\mu + A_{16}(\sigma^2/\mu) + A_{17}(\mu p/h) \\ & + A_{18}[(\sigma^2/\mu)\ (p/h)] + \epsilon. \end{aligned}$$

We fit the model to the system of 576 items using stepwise least-squares regression. The following expression yields a coefficient of determination of 0.998:

$$(17)\qquad \begin{aligned} T_a = {} & 1.110\, hW - .001049\, pW + .3364\, Kr \\ & -.2234\, h + .3274\, hD + .4476\, h\sigma^2/\mu + .003062\, p\sigma^2/\mu . \end{aligned}$$

Expression (17) is within 1.9% of actual values of T, on the average, when used in the 576-item system. It is accurate to within 4% for 89% of the items and to within 8% for 99% of the items. Only four items produce errors in excess of 10%. These items have μ equal 2, σ^2 equal 18, L equal 0, and p/h equal 4 or 9.

Although the approximation appears to be accurate in most cases, it may be inaccurate for policies that have significantly suboptimal values of s and S. This is because the differences between (16) and (17) suggest that the economics of optimal policies are intrinsic to the approximation obtained. The robustness of (17) is discussed explicitly in Section 4.

Approximating Backlog Cost

Attempts at finding a simple, accurate approximation for backlog cost were unsuccessful. Expression (9) was used to construct a regression model similar to those described above. The result was a coefficient of determination of 0.44. The relative errors were very large, in some cases exceeding 100%, making them significant even when compared on an absolute basis with other components of total cost.

The next attempt was to employ the identity

$$B = T - H - R$$

and use (10), (13), and (17) in place of R, H, and T. This approximation has an average percentage error of 18%, with large absolute errors for many of the items.

In order to get a reasonably accurate approximation, it was necessary to form a regression model that included all the variables appearing in the models for R, H, and T. It was also necessary to fit this model separately for optimal and power approximation policies and for each of the four settings of unit backorder penalty cost. That is, the 576-item system was partitioned into 8 systems of 72 items, and 8 separate regressions analyses were performed. The resulting approximation has an average coefficient of determination of 0.998. As the high coefficient of determination indicates, the fits are good in terms of absolute errors, although there are relative errors in excess of 70% for items with large values of p/h. However, the approximation is a complicated expression involving ten coefficients in each of the 8 subsystems (80 coefficients in all, for the 576-item system). Also, since the approximation was fit separately for each setting of p/h, there is no explicit functional dependence on this parameter. The reader is referred to [2] for additional details.

Backlog cost has proven to be surprisingly difficult to approximate. We point out that among the operating characteristics listed in (2), backlog cost is the most sensitive to the tail of the demand distribution. It appears that an accurate specification of the demand distribution is required for a reasonably precise calculation of backlog cost.

4. COMPUTATIONAL EXPERIENCE

We test the quality of approximations (10), (13), (14), (15), and (17) by using them in a multi-item system with the parameter settings of Table 2. Note that all the numerical parameters have values not found in the 576-item system. Each parameter has one interpolated value

TABLE 2 — *A 64-Item System with New Parameter Settings*

Factor	Levels	Number of Levels
Demand distribution	Negative Binomial ($\sigma^2/\mu = 5$) Negative Binomial ($\sigma^2/\mu = 15$)	2
Mean demand	0.5, 7.0	2
Replenishment lead time	1, 6	2
Replenishment setup cost	16, 48	2
Unit penalty cost	49, 132	2
Unit holding cost	1	1
Policy	Optimal policy, power approximation policy	2

and one extrapolated value. A full factorial combination of the values is used, yielding 64 items. The system is a rather severe test of robustness since only two items have all parameters with values within the ranges used to derive the approximations. There are 10 items with one extrapolated parameter, 20 items with two extrapolated parameters, 20 with three extrapolations, 10 with four extrapolations, and 2 items with all five parameters extrapolated.

We compare actual values of H, P, R, and T for the 64 items with our analytic approximations. Backlog cost B is not considered because of the complexity of our approximation and the absence of an explicit dependence on unit penalty cost. The average percent deviations from actual values of H, P, R, and T are 1.6%, 0.2%, 1.4%, and 2.6%, respectively. The distributions of percent deviations are summarized in Table 3. Our approximations are quite accurate considering the wide range of parameters spanned by the system.

TABLE 3 — *Percentage Deviations of Approximations in a 64-Item System*

(Entries are the number of items with errors in the given range, with the cumulative percentage of items in the system in parentheses.)

Range of Deviation	Holding Cost	Backlog Protection	Replenishment Cost	Total Cost
[0%,2%)	48 (75%)	64 (100%)	48 (75%)	30 (47%)
[2%,4%)	6 (84%)		8 (88%)	22 (81%)
[4%,6%)	5 (92%)		0 (88%)	6 (91%)
[6%,8%)	3 (97%)		6 (97%)	4 (97%)
[8%,10%)	2 (100%)		2 (100%)	1 (98%)
[10%,12%)				1 (100%)

The holding cost approximation is extremely accurate for all cases with μ greater than 0.5 or σ^2/μ less than 15. All items with deviations greater than 4% have μ equal 0.5 and σ^2/μ equal 15. If we consider only the items with fewer than two parameters extrapolated, the average error is 0.4%.

The backlog protection approximation is excellent, with only one item having a deviation in excess of 0.7%.

Our approximation for replenishment cost is also robust. All items with deviations in excess of 4% have μ equal 0.5, σ^2/μ equal 15, and K/h equal 16. Items with fewer than two extrapolated parameters have an average error of 0.1%.

Low μ and high σ^2/μ are also sources of large errors for our total cost approximation. All items with deviations in excess of 4% have either μ equal 0.5 or σ^2/μ equal 15, or both. Items with fewer than two extrapolated parameters have an average deviation of 1.2%.

We commented in Section 3 that the approximation for total cost may be inaccurate for items with significantly suboptimal values for s and S. The remark is equally valid for the backlog protection expressions (14) and (15), since they are based on a theoretical result for optimal policies. This issue is of interest to the analyst who may have reason to use an (s,S) policy which is designed to satisfy criteria other than simply minimizing total cost. We now proceed to illustrate how the accuracy of the approximations is affected when nonoptimal values are used for s and S.

Consider the following system of items that are controlled with nonoptimal policies. We choose a base-case item with σ^2/μ equal 5, μ equal 9, L equal 2, p/h equal 49, and K/h equal 48. The optimal value of (s,S) for this item is (43,73). We now use this *policy* on items with *different* parameter values. The new parameters are obtained by increasing or decreasing each base-case parameter value, one at a time, yielding 10 items. The parameter values of the system are displayed in Table 4. For each item we compare the actual (exactly computed) and approximate values of H, P, R, and T.

TABLE 4 — *Percentage Errors of Approximation for Nonoptimal Policies*

		Percentage Errors of Approximations			
Changed Value		Holding Cost	Backlog Protection	Replenishment Cost	Total Cost
σ^2/μ =	4 (−20%)	.07%	−.6%	−.00%	6.0%
	6 (+20%)	.05%	.7%	.00%	−5.0%
μ =	7 (−22%)	.11%	−1.3%	−.03%	13.7%
	11 (+22%)	−.04%	2.9%	.03%	−22.2%
L =	1 (−50%)	−.04%	−1.6%	.00%	12.6%
	3 (+50%)	.02%	5.5%	.00%	−36.2%
p/h =	39 (−20%)	−.01%	−.5%	.00%	3.9%
	59 (+20%)	−.01%	.3%	.00%	−2.2%
K/h =	38 (−21%)	−.01%	.0%	.00%	4.0%
	58 (+21%)	−.01%	.0%	.00%	−2.2%
Average of Absolute Values		.04%	1.3%	.01%	10.8%

Observe in Table 4 that the approximations for holding cost and replenishment cost are very accurate, with average percentage deviations of 0.04% and 0.01%, respectively. The approximation for backlog protection is somewhat less accurate, with the largest errors occurring for large values of lead time and mean demand. The total cost approximation does not perform well in the system, deviating by an average of 10.8%. Thus, we conclude that the approximations for backlog protection and total cost should be used with caution for significantly nonptimal policies. An approach to reducing the errors might be gleaned from the pattern of deviations in Table 4. Notice that when each parameter is larger than in the base case, the approximation underestimates the total cost, and when the parameter is smaller than in the base case, the approximation overestimates the total cost. The reverse is true for backlog protection.

Finally, we consider the issue of how well the approximations perform when the demand parameters are not accurately specified. This issue is of interest in applied settings when the mean μ and variance σ^2 of demand are not known but, rather, are estimated using past data. We have found that the approximations are rather robust when subjected to perturbations of this type. That is, the relative errors of the operating characteristic approximations tend to be smaller than the relative errors in the demand parameters. Furthermore, the errors are nearly symmetric so that when the operating characteristics of several items are aggregated, the errors due to high values of demand parameters tend to cancel those due to low demand parameters.

As an illustration we consider two items controlled by power approximation policies, one having a mean demand μ of 4 and the other having μ equal to 12. The other parameters of the items are identical; demand has a negative binomial distribution with σ^2/μ equal to 5, the lead time L is 2, the setup cost K is 48, the unit backorder penalty cost p is 49, and the unit holding cost h is 1. We measure the stability of the operating characteristic approximations by substituting perturbed demand parameters μ' and σ' in place of the correct values μ and σ, and comparing the approximated values with exactly computed values. For each of the items, we evaluated the approximations when μ'/μ and σ'^2/σ^2 took the values 0.80, 0.90, 0.95, 1.00, 1.05, 1.10, and 1.20. All combinations of perturbed values were tested, yielding 49 cases for each item, or a total of 98 cases.

We summarize the results in Table 5, where average absolute values of relative errors are listed for several ranges of demand parameter perturbations. Notice that the backlog protection approximation is not listed in Table 5. This is because the approximation is not a function of the demand parameters and, therefore, displays no variation when they are changed. The replenishment cost approximation displays the least stability in Table 5, with an average deviation of 6.9% for the 98 items. Errors ranged up to 19.5% for individual cases with extremely perturbed demand parameters. The holding cost approximation is more robust, yielding an average deviation of 4.7% and a maximum deviation of 13.7%. The approximation for total cost, however, has an average error of only 3.9% and a maximum error of 10.0%.

TABLE 5 — *Percentage Errors of Approximation When Demand Parameters Are Incorrectly Specified*

Range for Demand Parameters		Number of Cases	Average Absolute Value of Percentage Errors		
μ'/μ	σ'^2/σ^2		Replenishment Cost	Holding Cost	Total Cost
1.0	1.0	2	0.04%	0.2%	1.3%
[.95,1.05]	[.95,1.05]	42	3.0%	2.0%	1.6%
[.90,1.10]	[.90,1.10]	70	4.2%	2.8%	2.3%
[.80,1.20]	[.80,1.20]	98	6.9%	4.7%	3.9%

We note that the data in Table 5 are measures of the accuracy of the approximations for individual cases. A measure which is perhaps of greater interest in an applied setting is the aggregate error over all 98 cases, which is less than 0.5% for each of the characteristics. That is, when the 98 approximated values are averaged and compared with the exact average value, the difference is less than 0.5%. This observation can be regarded as evidence that the approximations are relatively unbiased when the demand parameters are replaced with unbiased statistics.

5. CONCLUSIONS

We have derived approximations for replenishment cost (10), holding cost (13), backlog protection (14), (15), and total cost (17). The expressions are quite accurate and are much easier to compute than the exact expressions (2). Additional simplification of calculations could result from using a normal distribution function in lieu of the function G in (12). Then the six evaluations of G in (12) could be replaced by terms involving the standard normal distribution function, which requires only a simple table look-up. This possibility has not yet been investigated.

Despite the good fits obtained in (10), (13), (14), (15), and (17), we caution against their use in certain circumstances. The results of Section 4 have demonstrated that the approximations for backlog protection and total cost become less accurate when used for significantly nonoptimal policies. Although the approximations for replenishment cost and holding cost are quite accurate over the investigated range of parameter settings, we suspect that they might break down when used for very small values of $D = S - s$. This is because (3) is based on an asymptotic expression for the renewal function $M(D)$.

ACKNOWLEDGMENTS

This paper is based on material from the author's Ph.D. dissertation, written at Yale University. The author is pleased to acknowledge the guidance and encouragement of his advisor, Professor Harvey M. Wagner. The author also wishes to thank the anonymous referee whose careful reading of an earlier manuscript helped to improve this paper significantly.

REFERENCES

[1] Ehrhardt, R., "The Power Approximation for Computing (s,S) Inventory Policies," Management Science, *25*, 777-786 (1979).

[2] Ehrhardt, R., "Operating Characteristic Approximations for the Analysis of (s,S) Inventory Systems," Technical Report No. 12, School of Business Adminstration and Curriculum in Operations Research and Systems Analysis, The University of North Carolina at Chapel Hill, Chapel Hill, N.C. (1977).

[3] Iglehart, D., "Optimality of (s,S) Policies in the Infinite Horizon Dynamic Inventory Problem," Management Science, *9*, 259-267 (1963).

[4] Roberts, D., "Approximations to Optimal Policies in a Dynamic Inventory Model," *Studies in Applied Probability and Management Science*, edited by K. Arrow, S. Karlin, and H. Scarf, (Stanford University Press, Stanford, Calif., 1962).

[5] Smith, W., "Asymptotic Renewal Theorems," Proceedings of the Royal Society (Edinburgh), *A 64*, 9-48 (1954).

[6] Veinott, A. and H. Wagner, "Computing Optimal (s,S) Inventory Policies," Management Science, *11*, 525-552 (1965).

OPTIMAL ORDERING POLICIES WHEN ANTICIPATING PARAMETER CHANGES IN EOQ SYSTEMS

B. Lev and H. J. Weiss

Temple University
Philadelphia, Pennsylvania

A. L. Soyster

Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

The classical Economic Order Quantity Model requires the parameters of the model to be constant. Some EOQ models allow a single parameter to change with time. We consider EOQ systems in which one or more of the cost or demand parameters will change at some time in the future. The system we examine has two distinct advantages over previous models. One obvious advantage is that a change in any of the costs is likely to affect the demand rate and we allow for this. The second advantage is that often, the times that prices will rise are fairly well known by announcement or previous experience. We present the optimal ordering policy for these inventory systems with anticipated changes and a simple method for computing the optimal policy. For cases where the changes are in the distant future we present a myopic policy that yields costs which are near-optimal. In cases where the changes will occur in the relatively near future the optimal policy is significantly better than the myopic policy.

1. INTRODUCTION

The classical Economic Order Quantity (EOQ) inventory model has several basic assumptions that yield the elegant solution of ordering $Q^* = \sqrt{2\lambda K/h}$ where λ, K and h are the traditional inventory parameters of demand, setup and holding, respectively. The most basic assumption is that all of the parameters are constant. Several systems have been examined in which either the demand rate or the purchase price may vary with time. (see Goyal [4], Buzacott [3], Naddor [9], Resh, Friedman and Barbosa [10], Barbosa and Friedman [1] and Sivazlian [13].). In all of these papers the parameter changes are continuous with time and furthermore only one parameter is permitted to change. In this paper we consider EOQ models in which any or all of the parameters may change at some future point in time.

The system we examine has two distinct advantages over the previous models. One obvious advantage is that a change in any of the costs is likely to affect the demand rate and we allow for this. The second advantage is that often, the times that prices will rise are fairly well known by announcement or by previous experience. If prices have risen January 1, April 1 and

July 1, it is very reasonable to anticipate a price rise on October 1. Also, price changes are more likely to jump than to be continuous with time.

In Section 2 of this paper we develop the inventory model and determine the necessary conditions for a policy to be optimal. In addition, we present a simple method for computing the optimal policy. Furthermore, a by-product of this method is a myopic policy. The myopic policy works well when the horizon is large enough and the price or demand change is far enough in the future. In Section 3, we present computational results for several different sets of parameters.

2. THE STRUCTURE OF AN OPTIMAL POLICY

Consider a finite horizon of length T that is partitioned into two disjoint time periods; the closed interval $[0,S]$ called period 1 and the half open interval $(S,T]$ called period 2. The costs associated with period 1 are a per unit cost c_1, a holding cost rate h_1, for all items brought into stock during period 1 and a setup cost $K_1 > 0$ charged against each order placed during the period. For items brought into stock during period 2 the unit cost, holding cost rate and setup cost are c_2, h_2 and K_2, respectively. Thus, S is a time at which any or all of the inventory costs may change. Also, the demand rate may change at S. Let λ_1 and λ_2 denote the demand rates during periods 1 and 2, respectively. A finite sequence of lot sizes is to be purchased to satisfy the demand. We assume that the initial inventory is zero, delivery is instantaneous, orders are placed only when the inventory level is zero and the discount factor is either ignored or included in the holding cost. The optimal policy for cases with a positive initial inventory is discussed later. Of course, if there are known lead times the results of this paper still hold but the orders are placed earlier according to the amount of the lead time.

The total cost, $Z(Q)$, for a single order of quantity Q with corresponding holding cost and purchase cost is $Z(Q) = K_i + h_i\, Q^2/2\lambda_i + c_iQ$. Theorem 1 limits the structure of the optimal policy as follows:

THEOREM 1: An optimal policy must have the property that

(a) all orders placed and depleted in period 1 are of the same size

and (b) all orders placed and depleted in period 2 are of the same size.

PROOF: Suppose Q_1 and Q_2 are the sizes of two consecutive orders placed and depleted in either period and let $Q = Q_1 + Q_2$.

The total cost of these two orders $Z(Q_1)$ as a function of Q_1 is given by

$$\begin{aligned} Z(Q_1) &= 2K_i + h_i[(Q_1)^2 + (Q_2)^2]/2\lambda_i + c_i(Q_i + Q_2) \\ &= 2K_i + h_i[Q_1^2 + (Q - Q_1)^2]/2\lambda_i + c_iQ. \end{aligned}$$

We have that the first and second derivatives are

$$Z'(Q_1) = h_i[2Q_1 - 2(Q - Q_1)]/2\lambda_i,$$

and

$$Z''(Q_1) = 4h_i/2\lambda_i > 0.$$

Hence, Z is strictly convex in Q_1 and is minimized only at $Q_1 = \dfrac{Q}{2} = Q_2$. Thus, two consecutive orders placed and depleted during the same period must be the same size, which implies that all orders placed and depleted in either one of these two periods must be the same size, and the theorem is proved.

Since the orders must be placed and depleted during the same period, Theorem 1 does not apply to an order that is placed on or before S (period 1) but depleted after S (period 2). Such an order is called a crossing order. Theorem 1 implies that the structure of the optimal ordering policy has been reduced to one of two possible forms depending on the inventory level at time S. If the inventory level is zero at S (Figure 1a), then the structure of the optimal policy is to place $m > 0$ orders of size $Q_1 = \lambda_1 S/m$ during $[0,S)$, place an order of size Q_a, $0 < Q_a \leqslant \lambda_2(T - S)$ at S, and place $n \geqslant 0$ orders of size $Q_2 = (\lambda_2(T - S) - Q_a)/n$ during period 2. (Note that if $n = 0$ then Q_2 does not exist). This case is denoted as the zero inventory case (ZIC). If the inventory level is positive at S (Figure 1b), then the structure of the optimal policy is to place $m \geqslant 0$ orders of size Q_1 before S, one order of size Q_a that crosses S and $n \geqslant 0$ orders of size Q_2 after S. This case is denoted as the nonzero inventory case (NZIC) and the two cases are examined separately.

2.1 Zero Inventory Case

The optimal number of orders to place for the finite horizon inventory model with parameters λ, h, K, T is given by Schwarz [12] as the integer n satisfying

$$n(n - 1) \leqslant h\lambda T^2/2K \leqslant n(n + 1). \tag{1}$$

The right hand inequality is

$$n^2 + n - h\lambda T^2/2K \geqslant 0.$$

The solution for the quadratic inequality is

$$n \geqslant - 1/2 + \sqrt{1/4 + h\lambda T^2/2K}.$$

The left hand inequality yields

$$n \leqslant 1/2 + \sqrt{1/4 + h\lambda T^2/2K}.$$

Since n is a positive integer

$$n = < - 1/2 + \sqrt{1/4 + h\lambda T^2/2K} >$$

where $< x >$ represents the least integer greater than or equal to x. Define an integer valued function $N(\lambda, h, K, T)$ of the inventory parameters as

$$N(\lambda, h, K, T) = < - 1/2 + \sqrt{1/4 + h\lambda T^2/2K} >. \tag{2}$$

(N is used if the parameters are clearly defined).

It follows that for the ZIC the optimal number of orders to be placed during $[0,S)$ is given by

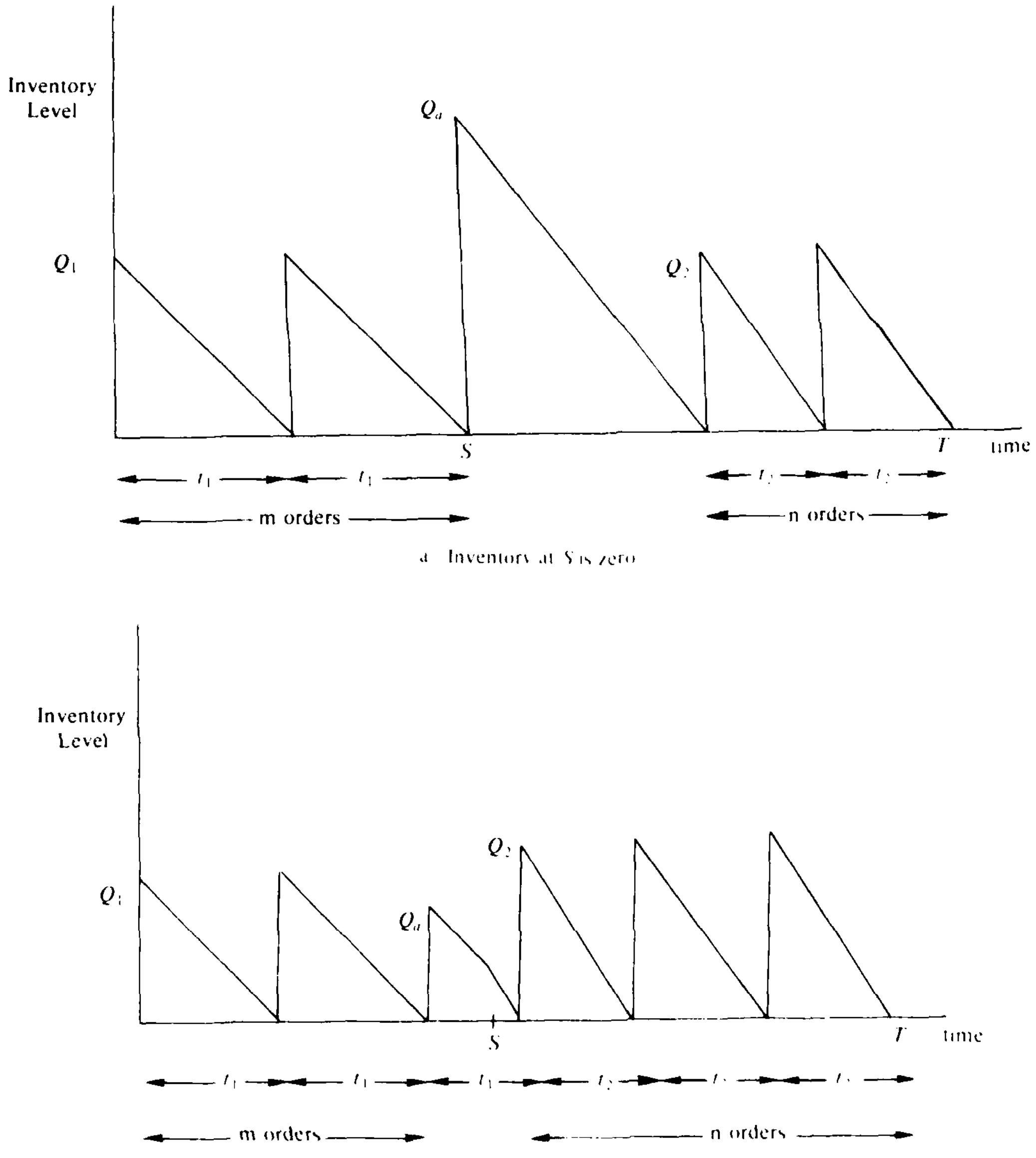
$$m^* = N(\lambda_1, h_1, K_1, S)$$

and the optimal order size is given by

$$Q_1 = \lambda_1 S/m^*$$

The costs incurred in $[S,T]$ are given by

$$f(t, n) = K_1 + h_1\lambda_2 t^2/2 + \lambda_2 t c_1 + n(K_2 + h_2\lambda_2 t_2^2/2 + \lambda_2 t_2 c_2)$$



a. Inventory at S is zero

b. Inventory at S is positive

FIGURE 1. Optimal Policy Structure-ZIC and NZIC

where t is the length of time it takes to deplete the order placed at S and $t_2 = (T - S - t)/n$. Letting $R = T - S$, $F(t, n)$ can be expressed as $F(t, n) = K_1 + h_1\lambda_2 t^2/2 + \lambda_2 t c_1 + nK_2 + h_2\lambda_2(R - t)^2/2n + \lambda_2(R - t)c_2$. The total ZIC costs are thus

$$F(t,n) + m^*K_1 + m^*h_1Q_1^2/2\lambda_1 + \lambda_1 Sc_1. \quad (3)$$

The partial derivative of (3) with respect to t, provides a necessary condition for (t,n) to minimize the total inventory costs for the zero inventory case:

$$0 = h_1\lambda_2 t + \lambda_2 c_1 - h_2\lambda_2(R - t)/n - \lambda_2 c_2 \quad (4)$$

or $t = (n(c_2 - c_1) + h_2R)/(nh_1 + h_2)$.

Notice that if the per unit cost increases then t will be positive. If the cost decreases then t may be negative. If this is the case then at time S an order should be placed for as few units as possible or the order should be delayed until time $S + \epsilon, \epsilon > 0$.

Also note that, if t is given by (4), then $R - t$ is the time in which the n orders are placed and is given by

$$\begin{aligned} R - t &= R - (n(c_2 - c_1) + h_2R)/(nh_1 + h_2) \\ &= n(h_1R - c_2 + c_1)/(nh_1 + h_2). \end{aligned}$$

This will be nonpositive if and only if $h_1R - c_2 + c_1$ is nonpositive. Thus, if $h_1R - c_2 + c_1 \leqslant 0$, then n must be zero and $t = R$. This means that if the cost of ordering one unit at price c_1 and incurring the holding cost h_1 for the entire span R is not more than c_2, the incremented purchase price, then obviously one should avoid any purchases at price c_2. If $h_1R - c_2 + c_1 > 0$, then $R - t$ is positive and $n \geqslant 1$. If $R - t$ is positive, then n must be the optimal number of orders for a finite horizon inventory model of length $(R - t)$. Let $I = \{1, 2, \ldots\}$ and $n^*(R - t)$ represent the optimal number of orders to place in the second period. Then from Equation (1)

$$\begin{aligned} n^*(R - t) &= \min\{n \in I{:}n(n + 1) \geqslant (\lambda_2h_2/2K_2)(R - t)^2\} \\ &= \min\{n \in I{:}n(n + 1) \geqslant (\lambda_2h_2/2K_2)(n(h_1R - c_2 + c_1)/(nh_1 + h_2))^2\} \\ &= \min\{n \in I{:}(n + 1)(nh_1 + h_2)^2/n \geqslant (\lambda_2h_2/2K_2)(h_1R - c_2 + c_1)^2\}. \qquad (5) \end{aligned}$$

One could compute n^* by sequentially searching the integers. However, there exists a more efficient scheme.

Consider the inequality given inside the braces in (5) expressed as an equality.

$$(nh_1 + h_2)^2 (n + 1)/n = (\lambda_2h_2/2K_2)(h_1R - (c_2 - c_1))^2.$$

Let

$$z = (\lambda_2h_2/2K_2)(h_1R - (c_2 - c_1))^2.$$

Then

$$(n^2h_1^2 + 2nh_1h_2 + h_2^2)(n + 1) - nz = 0$$

or

$$n^3h_1^2 + 2n^2h_1h_2 + nh_2^2 + n^2h_1^2 + 2nh_1h_2 + h_2^2 - nz = 0$$

or

$$n^3h_1^2 + n^2(2h_1h_2 + h_1^2) + \text{[illegible]} + 2h_1h_2 - z) + h_2^2 = 0.$$

This is a cubic equation and the three roots to the equation can be found using standard algebraic techniques (see, for example, Burington [2]). The cubic equation might have a single real root n_1 or three real roots $n_1, n_2, n_3, n_1 \leqslant n_2 \leqslant n_3$. In the former case, the solution to (5) is $n^* = <n_1>$, and in the latter case, the solution to (5) is

$$n^* = \begin{cases} <n_1> & \text{if } <n_1> \leqslant n_2 \\ <n_3> & \text{if } <n_1> > n_2. \end{cases}$$

Hence, (5) is easily solvable.

Since the zero inventory case is relatively easy to compute and often performs well as will be seen in the next section we refer to it as the myopic policy.

2.2 Nonzero Inventory Case

Define $t_1 = Q_1/\lambda_1$, $t_2 = Q_2/\lambda_2$ and t_a as the depletion times of the orders placed during periods 1, 2 and the crossing order respectively. Let m and n be the number of orders placed during periods 1 and 2 respectively. The total cost is given by

$$\begin{aligned}(6)\qquad F(n, m, t_1, t_2, t_a) = {} & m(K_1 + h_1\lambda_1 t_1^2/2 + c_1\lambda_1 t_1) + n(K_2 + h_2\lambda_2 t_2^2/2 + c_2\lambda_2 t_2)\\ & + K_1 + h_1\{\lambda_1(S - mt_1)^2/2 + (S - mt_1)\lambda_2(mt_1 + t_a - S) + (mt_1 + t_a - S)^2\lambda_2/2\}\\ & + c_1\lambda_1(S - mt_1) + c_1\lambda_2(mt_1 + t_a - S).\end{aligned}$$

Thus, the mathematical programming problem is:

$$\begin{array}{lll} & \text{minimize} & F(n, m, t_1, t_2, t_a)\\ (7) & \text{subject to} & mt_1 < S\\ (8) & & mt_1 + t_a > S\\ (9) & & mt_1 + t_a + nt_2 = T\\ & & n, m, t_1, t_2, t_a \geqslant 0\\ & & n, m \text{ integers.}\end{array}$$

Notice that due to constraint (9) the problem for a fixed m and n is a two-dimensional problem as t_2 is determined by the rest of the variables. The problem is still too difficult to approach as a mathematical programming problem because of the strict inequalities, so we reduce it to a one-dimensional problem with the following result.

THEOREM 2: For fixed m, n either $Q_1 = Q_a$ or ZIC is better than NZIC.

PROOF: The proof first shows that when m orders of size Q_1 are followed by a crossing order of size Q_a then it must be true that $Q_a = Q_1$. Let $R \geqslant S$ be the time at which Q_a is depleted and consider R as fixed. For constants m and R the relationship between Q_a and Q_1 is

$$(10)\qquad Q_a = (S - mQ_1/\lambda_1)\lambda_1 + (R - S)\lambda_2.$$

The order, holding and purchasing cost Z for the period $[0, R)$ as a function of Q_1 is

$$\begin{aligned}Z(Q_1) = {} & h_1[mQ_1^2/2\lambda_1 + (S - mQ_1/\lambda_1)\lambda_2(R - S)\\ & + \lambda_1(S - mQ_1/\lambda_1)^2/2 + \lambda_2(R - S)^2/2] + (m + 1)K_1 + c_1(\lambda_1 S + \lambda_2(R - S)).\end{aligned}$$

The function is minimized when the first derivative is zero or when

$$(11)\qquad mQ_1/\lambda_1 - m\lambda_2(R - S)/\lambda_1 - m(S - mQ_1/\lambda_1) = 0.$$

Notice that the second derivative is $(m + m^2)/\lambda_1 > 0$ since $m > 0$. Rearranging (11) yields

$$(12)\qquad Q_1 = (S - mQ_1/\lambda_1)\lambda_1 + (R - S)\lambda_2.$$

This Q_1 is the unique optimal order quantity and is equal to Q_a from (10) hence all orders are of the same size.

The decision variable Q_1 must satisfy the constraint $mQ_1/\lambda_1 \leqslant S$. If (12) violates this constraint the solution is on the boundary, i.e., $Q_1 = \lambda_1 S/m$ which means that all m orders placed strictly before S are of the same size and the theorem is proved.

We have that

$$Q_a = \lambda_1(S - mt_1) + \lambda_2(mt_1 + t_a - S)$$

and from Theorem 2 that $Q_a = Q_1 = \lambda_1 t_1$. Hence, it follows that

$$t_a = (\lambda_1 t_1 + (\lambda_2 - \lambda_1)(S - mt_1))/\lambda_2. \tag{13}$$

Furthermore, constraint (8) must be satisfied. Recall that using (13) and (8) one gets

$$\begin{aligned} mt_1 + t_a &= mt_1 + (\lambda_1 t_1 + (\lambda_2 - \lambda_1)(S - mt_1))/\lambda_2 \\ &= [(m + 1)\lambda_1 t_1 + (\lambda_2 - \lambda_1)S]/\lambda_2. \end{aligned} \tag{14}$$

Notice that $mt_1 + t_a > S$ if and only if $(m + 1)t_1 > S$. Thus, express $F(n, m, t_1, t_2, t_a)$ as a function of only one depletion time by substituting (13) and $t_2 = (T - (mt_1 + t_a))/n$ into the expression for $F(n, m, t_1, t_2, t_a)$ given by (6). Denote by $f(t_1)$ the cost for a fixed m and n when the depletion time is t_1. Then, after substitution

$$\begin{aligned} f(t_1) = {} & (m + 1)K_1 + nK_2 + \frac{mh_1\lambda_1 t_1^2}{2} + \frac{h_2}{2n\lambda_2}[\lambda_2 T - (m + 1)\lambda_1 t_1 - (\lambda_2 - \lambda_1)S]^2 \\ & + \frac{h_1\lambda_1(S - mt_1)^2}{2} + h_1(S - mt_1)\lambda_1[(m + 1)t_1 - S] + \frac{h_1\lambda_1^2}{2\lambda_2}[(m + 1)t_1 - S]^2 \\ & + c_1\lambda_1 t_1(m + 1) + c_2\lambda_2 T - (m + 1)c_2\lambda_1 t_1 - c_2(\lambda_2 - \lambda_1)S. \end{aligned} \tag{15}$$

Now $f'(t_1)$ is given by

$$\begin{aligned} f'(t_1) = {} & h_1\lambda_1 S - \frac{h_2}{n\lambda_2}[\lambda_2 T - (m + 1)\lambda_1 t_1 - {}'\lambda_2 - \lambda_1)S]\lambda_1(m + 1) \\ & - h_1\lambda_1[(m + 1)t_1 - S]m + \frac{h_1\lambda_1^2}{\lambda_2}[(m + 1)t_1 - S](m - 1) + \lambda_1(m + 1)(c_1 - c_2). \end{aligned} \tag{16}$$

Also,

$$f''(t_1) = (m + 1)^2\lambda_1^2[h_2/n - h_1]/\lambda_2 - h_1\lambda_1 m(m + 1). \tag{17}$$

Now if (17) is negative $f(\cdot)$ is concave and hence the minimum occurs at an extreme point of the feasible region. Thus, either the minimum is a zero inventory case or $t = T/(n + m + 1)$. If (17) is positive $f(\cdot)$ is convex and either the minimum is at an extreme point and again we have the zero inventory case or $t = T/(n + m + 1)$ or the minimum occurs by setting the derivative equal to zero. This leads to the following:

THEOREM 3: If for a fixed m and n the optimal case is the nonzero inventory case then either $t_1 = T/(m + n + 1)$ or

$$t_1 = \frac{h_2\lambda_2 T - (h_2 + nh_1)(\lambda_2 - \lambda_1)S + n\lambda_2(c_2 - c_1)}{(m + 1)\lambda_1(h_2 + nh_1) - nmh_1\lambda_2}. \tag{18}$$

Notice that if there are no changes then $t_1 = T/k$ where k is the number of orders that are of the same size as previously shown by Schwarz [12]. Also, if only the demand changes then $t_a = t_2$. Given t_1, the last task is to find m and n. As before, if t_1 is the depletion time of each of the first m orders, then $T - mt_1 - t_a$ is the length of time for the last n orders and the optimal number of orders placed during $[mt_1 + t_a, T]$ must satisfy Equation (1). That is, $n^*(T - mt_1 - t_a) = \min\{n \in I : n(n + 1) \geqslant (\lambda_2 h_2/2k_2)(T - mt_1 - t_a)^2\}$.

It appears that one needs to compute t_1 and n for all values of m. This would be a formidable task. However, the number of possible values for m can be reduced by the following:

THEOREM 4: For the case where the inventory level is positive at S either $m^* = N(S)$ or $m^* = N(S) + 1$ where $N(S)$ is the optimal number of orders to place in a finite horizon $[0, S]$. Furthermore, $n^* \geqslant N\left[T - \frac{m + 1}{m} S\right]$.

PROOF: Let $a = mt_1$. $S < a \leqslant T$. $a > S \rightarrow N(a) \geqslant N(S)$, since (2) is nondecreasing. All orders must be placed before S. Let b be the time of the last order. Then $N(b) \leqslant N(S)$, hence, $m^* \leqslant N(b) + 1$. Thus, either $N(S)$ or $N(S) + 1$ orders are placed. The restriction on n^* follows from Theorem 3 in [5].

We now can solve the NZIC for $m = N(S)$ and for $m = N(S) + 1$ and take the minimum cost of the ZIC and the NZIC. The algorithm is as follows:

1. Calculate $N(S)$ from (2) and set $m = N(S)$.
2. Calculate $N(T - S)$ from (2) and set $n = N(T - S)$.
3. Calculate $t^*(n)$ from (4) and compute the cost for the ZIC from (3)
4. For $n = N\left[T - \frac{m + 1}{m} S\right]$ to $N(T - S)$ calculate $t_1^*(m, n)$ from (18) and the cost for the NZIC from (15).
5. Set $m = N(S) + 1$.
6. Repeat step 4.
7. Find the minimum costs from steps 3,4,6

The last detail to discuss is that of an initial inventory. If the beginning inventory, I_0, is less that or equal to $\lambda_1 S$, obviously the inventory should be depleted and the problem is that of a finite horizon of length $T - I_0/\lambda_1$ with a price change at time $S - I_0/\lambda_1$. If the beginning inventory will not be depleted until after time S, obviously no purchases should be made until at least time S. In this case, the cost of not purchasing at S and then purchasing when the inventory is depleted should be compared with the cost of purchasing units at time S.

3. COMPUTATIONAL RESULTS

It is interesting to determine what effect varying the horizon or the time at which the parameters change would have on the optimal policy. In particular whether or not the myopic zero inventory case is optimal and if not how close to optimal it is. Note that for the case of no changes the optimal cost as a function of the horizon appears as in Figure 2 (see [5], [12]). Schwarz [11] has shown that if the horizon is at least 5 EOQs worth then the optimal finite horizon cost is no more than 1% above the optimal infinite horizon cost. One expects similar behavior in this model.

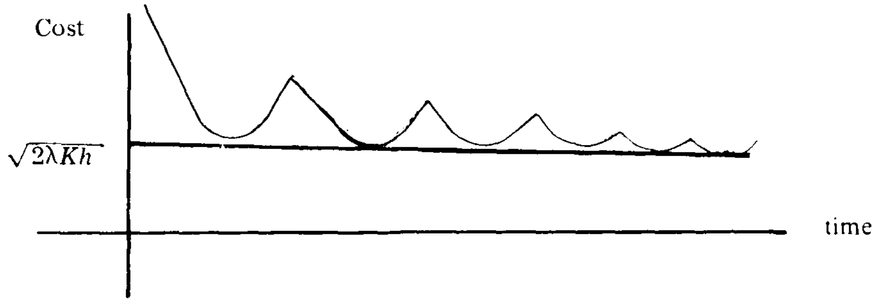


FIGURE 2 Optimal cost as a function of time when parameters remain constant

Table 1 contains the optimal costs for both the zero inventory case and nonzero inventory case where all parameters are fixed except for the horizon. The per unit cost was changed by .1 and the holding cost by .025. The demand and setup cost are constant throughout the two periods. Notice from Table 1 that the optimal policy alternates back and forth between the myopic and nonmyopic policies. Also, as the horizon becomes large the overcost when using the myopic policy tends to decrease. In fact, for any horizon above 25 the overcost is less than 1%. Incidentally, the infinite horizon optimal policy is the zero inventory case, with an average cost of 50.75.

TABLE 1 — *Inventory Costs as a Function of Horizon Length for* $\Delta C = .1$ *(2%) and* $\Delta h = .025$ *(2%)*

$\lambda = 5,\ K_1 = K_2 = 50,\ h_1 = 1.25,\ h_2 = 1.275,\ c_1 = 5,\ c_2 = 5.1,\ S = 20$			
T	ZIC Average Cost	NZIC Average Cost	ZIC/NZIC-1
21	53.66	50.02	7.28%
22	52.58	50.09	4.97
23	51.74	50.02	3.44
24	51.09	50.00	2.18
25	50.63	50.14	.98
26	50.32	50.17	.30
27	50.15	50.12	.06
28	50.11	50.40	—
29	50.17	50.30	—
30	50.34	50.23	.22
31	50.60	50.19	.82
32	50.19	50.41	—
33	50.23	50.33	—
34	50.34	50.29	.10
35	50.26	50.25	.02
36	50.25	50.42	—
37	50.28	50.36	—
38	50.36	50.32	.08
39	50.30	50.30	—
40	50.30	50.44	—

Table 2 contains similar information but for a larger price increase. Let $\Delta C = 1$ and $\Delta h = .25$ while all other parameters are as above. Again, when the horizon is 25 or larger the myopic policy is never worse than 1% above optimal. However, in this case the myopic policy is optimal for all horizons larger than 35.

TABLE 2 — *Inventory Costs as a Function of Horizon Length for* $\Delta C = 1$ *(20%) and* $\Delta h = .25$ *(20%)*

$\lambda = 5,\ K_1 = K_2 = 50,\ h_1 = 1.25,\ h_2 = 1.5,\ C_1 = 5,\ C_2 = 6,\ S = 20$			
T	ZIC Average Cost	NZIC Average Cost	ZIC/NZIC-1
21	53.71	50.02	7.38%
22	52.75	50.09	5.31
23	52.01	50.02	3.98
24	51.48	50.00	2.96
25	51.12	50.63	.97
26	50.93	50.90	.06
27	50.87	51.23	—
28	50.94	51.38	—
29	51.12	51.61	—
30	51.41	51.90	—
31	51.79	51.99	—
32	52.24	52.11	.25
33	51.89	52.44	—
34	52.12	52.50	—
35	52.41	52.58	—
36	52.34	52.89	—
37	52.50	52.93	—
38	52.69	52.99	—
39	52.74	53.06	—
40	52.85	53.30	—

In the examples presented in Table 3 the horizon is fixed and the time of price change varies. The remaining parameters are identical to those of Table 1. The Table also contains which case is optimal in the long run. Notice how in the infinite horizon model as in the finite horizon model the cases alternate as S changes. Also, as S approaches T the myopic policy worsens.

In the next example presented in Table 4, S varies, and we use the larger cost increase as in Table 2. This time, the infinite horizon models always are optimized by the myopic policy. Again, as S approaches T the myopic policy begins to worsen.

The last set of examples given in Table 5 indicates that as the number of orders (using either policy) increases then the difference between the myopic and optimal policies lessens. The data used to generate Table 5 is identical to the data for Table 1 except that the holding cost is reduced from 25% of the purchase cost to 5% of the purchase cost. Notice that this generates fewer orders which in turn increases the overcost.

TABLE 3 — *Inventory Costs as a Function of the Time of Price Changes for* $\Delta C = .1$ *(2%) and* $\Delta h = .025$ *(2%)*

$\lambda = 5,\ K_1 = K_2 = 50,\ h_1 = 1.25,\ h_2 = 1.275,\ c_1 = 5,\ c_2 = 5.1,\ T = 30$
(and $T = \infty$ for last column)

S	ZIC	NZIC		ZIC/NZIC-1	$T = \infty$ Optimal Case
6	50.71	50.60	2,5*	.22%	NZIC
7	50.54	50.60		—	ZIC
8	50.54	50.50	3,5	.06	ZIC
9	50.50	50.50	3,4	—	NZIC
10	50.54	50.50		.08	NZIC
11	50.43	50.50		—	ZIC
12	50.45	50.41	4,4	.04	ZIC
13	50.38	50.41		—	NZIC
14	50.40	50.38	4,3	.04	NZIC
15	50.33	50.38		—	ZIC
16	50.38	50.32	5,3	.12	ZIC
17	50.28	50.32		—	NZIC
18	50.28	50.27	5,2	.02	NZIC
19	50.64	50.27		.74	ZIC
20	50.34	50.23	6,2	.22	ZIC
21	50.19	50.23		—	NZIC
22	50.17	50.16	6,1	.02	NZIC
23	50.15	50.14	7,1	.02	ZIC
24	50.28	50.14		.28	ZIC
25	50.54	50.14		.80	NZIC

The notations should be read as follow:
*The optimal policy for $S = 6$ and $S = 7$ is $m = 2$ $n = 5$

TABLE 4 — Inventory Costs as a Function of the Price Changes
for $\Delta C = 1$ (20%) and $\Delta h = .25$ (20%)

$\lambda = 5,\ K_1 = K_2 = 50,\ h_1 = 1.25,\ h_2 = 1.5,\ c_1 = 5,\ c_2 = 6,\ T = 30$ (and $T = \infty$)				
S	ZIC	NZIC	ZIC/NZIC-1	$T = \infty$
6	55.00	55.16	—	ZIC
7	54.59	55.16	—	ZIC
8	54.50	55.16	—	ZIC
9	54.13	54.71	—	ZIC
10	53.93	54.20	—	ZIC
11	53.60	54.00	—	ZIC
12	53.40	54.22	—	ZIC
13	53.11	54.22	—	ZIC
14	52.91	53.26	—	ZIC
15	52.84	52.91	—	ZIC
16	52.41	52.91	—	ZIC
17	52.11	52.93	—	ZIC
18	52.48	52.38	—	ZIC
19	51.87	51.89	—	ZIC
20	51.41	51.56	—	ZIC
21	51.11	51.56	—	ZIC
22	50.95	50.94	—	ZIC
23	50.80	50.94	—	ZIC
24	50.80	50.78	.04%	ZIC
25	50.95	50.63	.64	ZIC

TABLE 5 — *Inventory Costs as a Function of Horizon Length*
for $\Delta C = .1$ (2%) and $\Delta h = .005$ (2%)

$\lambda = 5,\ K_1 = K_2 = 50,\ h_1 = .25,\ h_2 = .255,\ c_1 = 5,\ c_2 = 5.1,\ S = 20$			
T	ZIC Average Cost	NZIC Average Cost	ZIC/NZIC-1
21	40.50	36.32	11.50%
22	39.85	36.40	9.47
23	39.28	36.31	8.18
24	38.79	36.25	7.01
25	38.36	36.20	5.95
26	37.99	36.18	4.99
27	37.67	36.18	4.12
28	37.40	36.19	3.34
29	37.16	36.21	2.62
30	36.97	36.25	1.98
31	36.80	36.50	.99
32	36.66	36.39	.74
33	36.56	36.36	.54
34	36.47	36.34	.37
35	36.42	36.33	.24
36	36.38	36.32	.14
37	36.36	36.33	.07
38	36.36	36.35	.02
39	36.37	36.37	< .01
40	36.40	36.48	—

In summary, if the horizon is large, compared with the time of price change (we suspect that large is 5 EOQs) then the myopic policy appears to be very worthwhile.

ACKNOWLEDGMENT

We thank the anonymous referee for his suggestions and corrections.

BIBLIOGRAPHY

[1] Barbosa, L.C. and M. Friedman, "Deterministic Inventory Lot Size Models—A General Root Law," Management Science *24*, 819-826 (1978).

[2] Burington, R.S., *Handbook of Mathematical Tables and Formulas*, 4th Edition (McGraw-Hill, New York, N.Y. 1965).

[3] Buzacott, J.A., "Economic Order Quantity with Inflation," Operational Research Quarterly, *26*, 3 (1975).

[4] Goyal, S.K., "An Inventory Model for a Product for which Purchase Price Fluctuates," New Zealand *Operational* Research, *3*, 2 (1975).

[5] Lev, B. and A.L. Soyster, "Inventory Models with Finite Horizons and Price Changes," Operational Research Quarterly, *30*, 1, 43-53 (1979).

[6] Lev, B., H. J. Weiss and A.L. Soyster, "Comment on an Improved Procedure for the Finite Horizon and Price Changes Inventory Model," Operational Research Quarterly, *30*, 9, 840-842 (1979).

[7] Lippman, S.A., "Economic Order Quantities and Multiple Set Up Costs," Management Science, *18*, 39-47 (1971).

[8] Ludin, R.A. and T.E. Morton, "Planning Horizons for the Dynamic Lot Size Model: Zabel vs. Protective Procedures and Computational Results," Operations Research, *23*, 711-734 (1975).

[9] Naddor, E., *Inventory Systems*, 48-50 (John Wiley and Sons, New York, N.Y. 1966).

[10] Resh, M., M. Friedman and L.C. Barbosa, "On a General Solution of the Deterministic Lot Size Problem with Time Proportional Demand," Operations Research, *24*, 718-725 (1976).

[11] Schwarz, L.B., "A Note on the Near Optimality of '5-EOQ's Worth' Forecast Horizons," Operations Research, *25*, 533-536 (1977).

[12] Schwarz, L.B., "Economic Order Quantities for Products with Finite Demand Horizons," AIIE Transactions *4*, 234-237 (1972).

[13] Sivazlian, B.D. and L.E. Stanfel, *Analysis of Systems in Operations Research* (Chapter [illegible]) (Prentice Hall, Englewood Cliffs, N.J. 1975).

SYSTEMS DEFENSE GAMES: COLONEL BLOTTO, COMMAND AND CONTROL*

Martin Shubik

Yale University
New Haven, Connecticut

Robert James Weber

Northwestern University
Evanston, Illinois

ABSTRACT

The classical "Colonel Blotto" games of force allocation are generalized to include situations in which there are complementarities among he targets being defended. The complementarities are represented by means of a system "characteristic function," and a valuation technique from the theory of cooperative games is seen to indicate the optimal allocations of defense and attack forces. Cost trade-offs between systems defense and alternative measures such as the hardening of targets are discussed, and a simple example is analyzed in order to indicate the potential of this approach.

1. COLONEL BLOTTO GAMES

The first example of what has come to be called a "Colonel Blotto game" was apparently given by Borel [3]. He discussed the case of a defender attempting to protect several locations against an aggressor. A typical objective of the aggressor was to maximize the expected number of locations captured.

Games involving this type of objective were subsequently studied by Tukey [11] and others (for example, Gross [7], Blackett [2], Dresher [4], Beale and Heselden [1]). As defined by Beale and Heselden, a (Colonel) *Blotto game* is a zero-sum game involving two opposing players, I and II, and n independent battlefields. I has A units of force to distribute among the battlefields, and II has B units. Each player must distribute his forces without knowing his opponent's distribution. If I sends x_k units and II sends y_k units to the kth battlefield, there is a payoff $P_k(x_k, y_k)$ to I as a result of the ensuing battle; the payoff for the game as a whole is the sum of the payoffs at the individual battlefields.

In this paper we consider a generalization of the classical Blotto game. This generalization gives regard to the important class of military problems wherein there exist complementaries among the points being defended. In such cases, the final status of the competitors is not determined merely by totalling individual target values, but depends on the relative value of

*This research was supported in part by a contract with the U. S. Office of Naval Research.

capturing (or neutralizing) various configurations of targets. Our generalization includes the classical Blotto games, as well as, for example, games in which the aggressor's objective is to maximize the probability of capturing a majority of the targets.

By considering complementarities among targets, we are in a position to study the defense of networks. For the purposes of increased reliability and security, redundancy is often intentionally incorporated into telephone and electrical power grids, early warning networks, and command and control systems. It is natural to ask how well protected such systems are from a disabling attack. Furthermore, it is of interest to consider cost trade-offs between built-in redundancy and extrinsic defense. In order to pursue these issues, we first introduce some terminology from cooperative game theory.

2. SYSTEMS PERFORMANCE AND THE CHARACTERISTIC FUNCTION

An n-person game in coalitional form is described by a *characteristic function* $v(\cdot)$ defined for all subsets of the set N of "players." When one is considering networks (or battlefields, or strategically important facilities), $v(S)$ may be interpreted as the value remaining in the system if only the set of nodes S is held. The characteristic function captures in a general setting the many types of complementarity which can exist among the various combinations of points in the network. (In traditional cooperative game theory it is frequently assumed that the characteristic function is superadditive; that is, if S and T are disjoint then $v(S) + v(T) \leq v(S \cup T)$. However, in the context of strategic systems this assumption may not be reasonable. If one is protecting a network of doomsday devices, for example, the characteristic function might assign a value of 1 to every nonempty set.)

There are many different "solutions" which have been suggested by game theorists for games in coalitional form. They reflect various aspects of the cooperative dealings among players with different goals. We note in particular the value solutions, which can be given an interpretation in terms of the military problem of allocating forces to a system of n nodes. In order to give this interpretation in detail we must reformulate the original n-person game as a two-person noncooperative game.

3. THE NONCOOPERATIVE GAME

We recast the given game as if it were a zero-sum game played between two opponents, a defender and an attacker. The n players in the original game are regarded as nodes (or individual targets) in a strategic network that the defender is trying to protect and the attacker is trying to destroy.

Let A and B be the respective amounts of strategic resources (troops, for example, or antiballistic and ballistic missiles) held by the defender and the attacker. The defender may choose any nonnegative allocation $x = (x_1, \ldots, x_n)$ of resources, subject to the constraint that $\Sigma x_i = A$. Similarly, the attacker may choose any allocation $y = (y_1, \ldots, y_n)$ for which $\Sigma y_i = B$. Let $f_i(x_i, y_i)$ be the function (yet to be specified) which indicates the outcome of the battle at point i. A natural interpretation which we take at this time is that $f_i(x_i, y_i)$ is the probability that the defender retains point i.

Assume that the goal of the defender is to maximize the (expected) effectiveness of the surviving configuration of targets. If the interests of the attacker are directly opposed to those of the defender, then we have at hand a two-person zero-sum game. The probability that the targets in the set S survive, while all others are destroyed, is

$$\prod_{i \in S} f_i(x_i,y_i) \prod_{i \notin S} (1 - f_i(x_i,y_i)).$$

Therefore, the expected effectiveness of the surviving collection is

$$\sum_{S \subseteq N} \left\{ \prod_{i \in S} f_i(x_i,y_i) \prod_{i \notin S} (1 - f_i(x_i,y_i)) \right\} v(S) ;$$

this is the defender's payoff.

If we suspend the interpretation of the functions f_i as probabilities, we find that this competitive game is indeed a direct generalization of the traditional Colonel Blotto game. Assume that the underlying characteristic function is additive, so that $v(S) = \sum_{k \in S} v(\{k\})$ for all $S \subseteq N$.

Then

$$D(x,y) = \sum_{k=1}^{n} f_k(x_k,y_k) \; v(\{k\}) .$$

By identifying $P_k(x_k,y_k)$ with $f_k(x_k,y_k) \cdot v(\{k\})$ (for example, by taking $P_k = f_k$ and $v(\{k\}) = 1$ for all $k \in N$), we can represent any desired classical Blotto game.

4. BATTLE MODELS

A listing of the various battle models which have been considered is beyond the scope of this paper. Moreover, a critical evaluation of the relative validity of these models does not appear to be available. Even Napoleon's dictum that God is on the side of the strongest battalion does not appear to be borne out when the force sizes of victors and losers in major battles are compared (for example, see Dupuy, page 89 [6]).

For the purposes of this paper we have chosen to consider a moderately general class of models in which the attacker and defender have homogenous resources. Hence, force mix problems have been set aside. Still, while it may be reasonable to assume that the probability that a target i is captured or destroyed is simply a function $f_i(x_i,y_i)$ of the resources expended in attack and defense by the two sides, the actual form of this function depends on empirical factors such as target type, physical vulnerability, troop morale, and the like.

We specifically consider outcome functions of the form

$$f(x,y) = \frac{\gamma x^m}{\gamma x^m + (1 - \gamma) y^m} ,$$

where we set $f(0,0) = \gamma$. The parameter γ may be interpreted as an indicator of the natural defensibility of the target; if $x = y$, then $f(x,y) = \gamma$. The homogeneity of the function f allows us to concern ourselves with the ratio $k = x/y$ of defending to attacking forces, rather than with the specific amounts x and y. The parameter m reflects the importance of the relative difference in size between the attacking and defending forces.

In the limit, as m becomes large, the outcome function becomes the crudest form of "superior forces" model the side which commits a greater force will win with certainty. If the resources of the defender and the attacker are of comparable size, in this limiting case the force-allocation game may fail to have a solution in pure strategies. (For an investigation of the degree of disparity of initial force sizes sufficient to guarantee the existence of optimal pure strategies, see Young, [13]).

On the other hand, if m is not too large, the outcome function is relatively insensitive to small changes in opposing allocations. We consider this case in the next section.

5. VALUE SOLUTIONS

Let v(·) be a characteristic function on N, and let $p = (p_1, \ldots, p_n)$ be a vector of probabilities (that is, each $0 \leqslant p_i \leqslant 1$). Then the $(p_1, \ldots, p_n)$-*value* of v is the n-vector $\beta = (\beta_1, \ldots, \beta_n)$ defined for all $i \in N$ by

$$\beta_i = \sum_{S \subseteq N - i} \left\{ \prod_{j \in S} p_j \prod_{\substack{k \in N - S \\ k \neq i}} (1 - p_k) \right\} [v(S \cup i) - v(S)].$$

Consider the force-allocation game based on v, in which the initial resources of the opposing sides are A and B, respectively. Assume that the outcome function at the kth target is defined by $f_k(x,y) = \gamma_k x^m / (\gamma_k x^m + (1 - \gamma_k) y^m)$. Then if both sides have optimal pure strategies, these strategies must be force allocations proportional to the $(f_1, \ldots, f_n)$ (A,B)-value of the underlying game. Furthermore, for all sufficiently small values of m, allocations proportional to the $(f_1, \ldots, f_n)$ (A,B)-value are indeed optimal.

Further details concerning these results are presented elsewhere (Shubik and Weber [9]).

6. THE COSTS OF SYSTEMS DEFENSE

"What price freedom?" is an important question, but one which political philosophers, economists, and Department of Defense budget proposers often find difficult to make precise. A model which links the value and cost of defense is presented here. (A different model is presented in Section 7, where we take the cost of defense as given but consider the possibility of trade-offs between direct defense and the physical reinforcement of individual targets.)

At an abstract level, there are four major items in the description of a defensive system: the military or societal "worth" of defense; the type, quantity, and structure of defensive forces; the cost of these forces; and the "hardness" (defensive strength) of individual targets.

The model of Section 3 avoids the problem of comparing value and cost by representing value within the characteristic function and taking as given the available attack and defense forces. Thus, constraints on military resources enter only as boundary conditions on a force assignment problem, rather than as a result of taking resource costs into account in the payoff structure.

We can modify the games of Section 3 to include costs in the following manner. The defender and attacker first select force levels k_1 and k_2, incurring costs of $c_1(k_1)$ and $c_2(k_2)$. They then each assign forces, and the payoffs are given by

$$(*) \qquad \begin{aligned} P_D &= v(S) - c_1(k_1), \text{ and} \\ P_A &= w(S') - c_2(k_2), \end{aligned}$$

where v(S) is the worth (in monetary units) to the defender of the configuration S of surviving targets, and $w(S')$ is the worth to the attacker of destroying or capturing the targets in S'. This is a two-stage nonconstant-sum game, which might be studied in terms of either equilibrium or minimax theories.

The fact that the above game formulates well as a two-stage process calls attention to the fact that the two stages are separate in both time and bureaucratic control. The problem for a defense department in dealing with the government as a whole is to select k_i, incurring the budgetary expense $c_i(k_i)$. The problem of the commander, having been presented with forces k_i, is to allocate these forces wisely.

From the viewpoint of analysis, the models of Section 3 seem worth pursuing at the level of command and control. However, it appears that the first stage of the model suggested by (*) concerns a very different aspect of decision making, and involves deep issues in the area of defense budgeting (some of these issues have been discussed by Hitch and McKean [8]).

7. THE HARDENING OF TARGETS

In order to illustrate some of the preceding considerations, we analyze a simple example. Assume that a defender seeks to protect three sites, at each of which several antiballistic missiles are siloed. If the attacker destroys any two (or all three) of the targets, the overall defensive system will collapse. The first site houses more missiles than the second, which in turn houses more than the third; although any two surviving sites will yield an adequate system, the survival of all three provides even greater security. We model this situation with a characteristic function v, which satisfies v(123) = 4; v(12) = 3; v(13) = 2; v(23) = 1; and v(S) = 0 if $|S| \leqslant 1$.

Assume that the attacker and defender possess comparable amounts of strategic resources, say, $A = B = 1$. Let the outcome of conflict at site k be represented by the function $f_k(x,y) = \gamma_k x/(\gamma_k x^m + (1-\gamma_k)y^m)$, for some moderately small value of m (that is, assume that equal forces engaged at site k will yield a result *favorable* to the defender with probability γ_k, and further assume that small differences in resource assignments lead to only relatively small changes in this probability). The parameter γ_k indicates the "hardness" of the target at site k (that is, its natural strength against attack). It follows, as was indicated in Section 5, that the optimal allocation of strategic forces by each side will be proportional to the $(\gamma_1, \gamma_2, \gamma_3)$-value of the game v. Hence, this allocation will be proportional to the vector

$$\beta = (2\gamma_2 + 3\gamma_3 - 2\gamma_2\gamma_3,\ 3\gamma_1 + \gamma_3 - 2\gamma_1\gamma_3,\ 2\gamma_1 + \gamma_2 - 2\gamma_1\gamma_2).$$

In particular, if we initially have $\gamma_1 = \gamma_2 = \gamma_3 = 1/2$, the optimal allocation for each side is (4/9, 3/9, 2/9).

Now, assume that additional capital is available to the defender, and may be used to harden any of the targets. Specifically, assume that an investment of Δc_k units of capital at site k will yield an increase of $(1 - \gamma_k)\Delta c_k$ in the hardness of target k, that is, $\partial\gamma_k/\partial c_k = (1 - \gamma_k)$. A natural question is how best to invest the additional capital.

Let the defender's allocation of forces be $x = (x_1, x_2, x_3)$, while the attacker's deployment is $y = (y_1, y_2, y_3)$. Then the value of the outcome of the competitive game, to the defender, is

$$D(x,y) = 3f_1f_2 + 2f_1f_3 + f_2f_3 - 2f_1f_2f_3,$$

where each f_k is evaluated at (x_k, y_k). The optimal strategies are $x^* = y^* = \beta/\Sigma\beta$. Therefore, the rate of gain from investment in the hardening of target k is

$$\frac{\partial D}{\partial c_k}(x^*,y^*) = \frac{\partial D}{\partial p_k}(x^*,y^*) \cdot \frac{\partial f_k}{\partial \gamma_k}(x^*,y^*) \cdot \frac{\partial \gamma_k}{\partial c_k}$$

$$= (\beta_k/\Sigma\beta_j) \cdot 1 \cdot (1 - \gamma_k).$$

The best investment is in the target (or targets) for which this expression is maximized. But the expression varies with the parameters γ_1, γ_2, and γ_3. Hence, if we begin with all γ_k equal, it is best to initially invest in work at the site for which β_k is maximal; this changes β as well as γ_k, after which we can determine the best target for further investment. Beginning with $\gamma_1 = \gamma_2 = \gamma_3 = 1/2$, we obtain the results indicated in the figures. (As the available capital increases without limit, the value of $D(x^*,y^*)$ approaches 4, and the three sites attract nearly equal proportions of the capital.)

This example illustrates several, but by no means all, of the types of computations which appear to be feasible and relevant to the study of tradeoffs in defense, in the hardening of targets, and in built-in system redundancy.

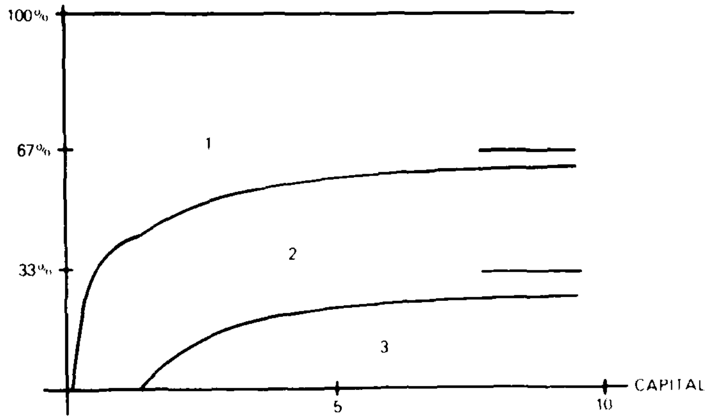


FIGURE 1 Allocation of capital to target reinforcement

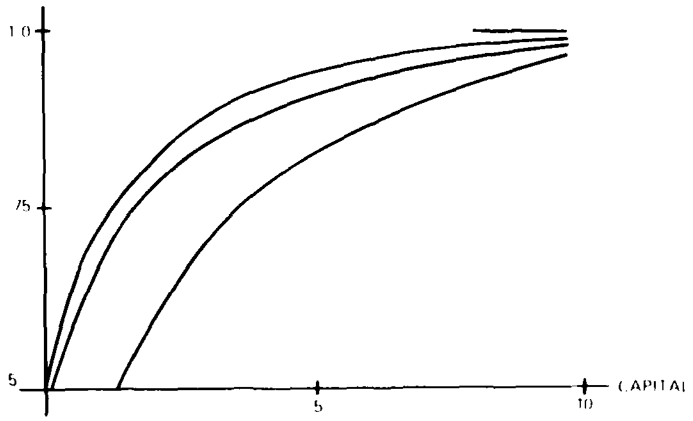


FIGURE 2 Hardness of targets x_1, x_2 and x_3

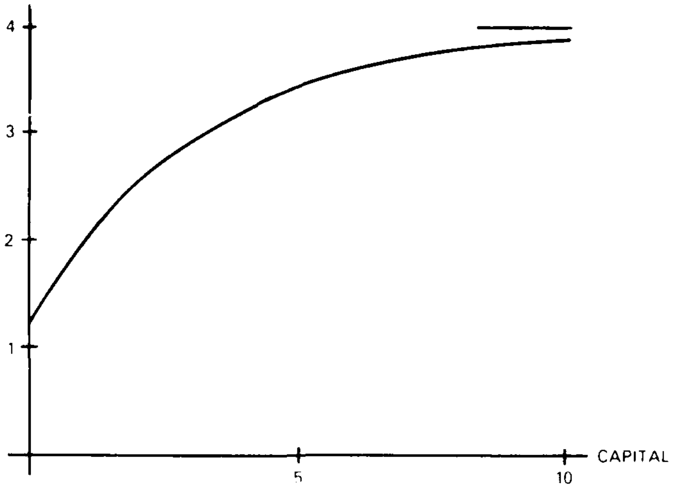


FIGURE 3. Value of game to defender $D(x, 1)$

BIBLIOGRAPHY

[1] Beale, E.M.L. and G.P.M. Heselden, "An Approximate Method of Solving Blotto Games," Naval Research Logistics Quarterly 9, 65-79 (1962).

[2] Blackett, D.W., "Pure Strategy Solutions to Blotto Games," Naval Research Logistics Quarterly 5, 107-109 (1958).

[3] Borel, E., *Traite du calcul des probabilites et des ses applications, Applications des jeux de hasard*, Vol. IV, Fascicule 2, (Gauthier-Villars, Paris, France, 1938).

[4] Dresher, M., *Games of Strategy: Theory and Applications*, (Prentice Hall, Englewood Cliffs, N.J., 1961).

[5] Dubey, P., A. Neyman and R.J. Weber, "Value Theory without Efficiency," Mathematics of Operations Research (to appear).

[6] Dupuy, T.N., "Analyzing Trends in Ground Combat," History, Numbers, and War, *1*, 2 79-91 (1977).

[7] Gross, O. and R. Wagner, "A Continuous Colonel Blotto Game," RAND Memorandum 408 (1950).

[8] Hitch, C.J. and R.M. McKean, *The Economics of Defense in the Nuclear Age* (Harvard University Press, Cambridge, Mass., 1960).

[9] Shubik, M. and R.J. Weber, "Competitive Valuation of Cooperative Games," Mathematics of Operations Research (to appear).

[10] Shubik, M. and H.P. Young, "The Nucleolus as a Noncooperative Game Solution," Cowles Foundation Discussion Paper No. 478, Yale University (1978).

[11] Tukey, J.W., "A Problem of Strategy," Econometrica *17*, 73 (1949).

[12] Weber, R.J., "Probabilistic Values for Games," Cowles Foundation Discussion Paper No. 471R, Yale University (1978).

[13] Young, H.P., "Power, Prices and Incomes in Voting Systems," International Institute for Applied Systems Analysis RR-77-5, March 1977.

ON NONPREEMPTIVE STRATEGIES IN STOCHASTIC SCHEDULING

K. D. Glazebrook

University of Newcastle upon Tyne
Newcastle upon Tyne, England

ABSTRACT

It is shown that there is an optimal strategy for a class of stochastic scheduling problems which is nonpreemptive. The results which yield this conclusion are generalizations of previous ones due to Glazebrook and Gittins. These new results also lead to an evaluation of the performance of nonpreemptive strategies in a large class of problems of practical interest.

1. INTRODUCTION

A job shop consists of one machine and a set $J = \{1, 2, \ldots, K\}$ of jobs to be processed on it. In general the processing time P_i for job i is a positive integer-valued random variable with known honest distribution, processing times for different jobs being independent. If job i is completed at time F_i (flow time) its cost is $C_i(F_i)$. There is a precedence relation R on the set J such that if $(i, j) \in R$ then the machine must complete job i before it can begin processing job j.

For simplicity, the major part of the material will be devoted to problems in discrete time. During each time interval $[t, t+1)$, $t \in Z^+ \cup \{0\}$, just one of the unfinished jobs is processed by the machine. A feasible strategy π is any rule for deciding how to choose the jobs in J for processing which is consistent with R. Under strategy π job i is completed at the random time $F_i(\pi)$. The objective is to find those strategies π in some given subset of the set of feasible strategies which minimize the total expected cost

$$TC(\pi) = E\left\{\sum_{i=1}^{K} C_i[F_i(\pi)]\right\}.$$

The economic criteria which have been most widely studied in this context are the discounted costs criterion, that is

$$C_i(F_i) = -K(i)a^{t}, \{0 < K(i)\}, (0 \leqslant a < 1), i \in J \tag{1}$$

(see [1], [2], [3], [4], [6], and [9]), and the criterion involving linear costs, that is

$$C_i(F_i) = K(i)F_i, \{0 < K(i)\}, i \in J \tag{2}$$

(see [2], [3], [6], and [10]).

The problem of finding optimal feasible permutations of J for the above economic criteria has essentially been solved in the sense that algorithms have been given which can be shown to generate all the optimal permutations. For details of this work see [4], [6] and [10]. Much work, however, remains to be done on the efficiency of these algorithms.

The problem of finding strategies which are optimal in the set of all feasible strategies for economic criteria (1) and (2) is much more difficult. Glazebrook [2] gave a characterization of the optimal strategies for the case when R has a digraph representation which is an out-tree. Results in a similar vein, though obtained in a rather different way, were reported by Meilijson and Weiss [5]. The problem with general R seems very complex.

Not surprisingly, then, concentration has latterly focused on the problem of giving a characterization of those problems which have an optimal strategy (in the set of all feasible strategies) given by a fixed permutation of J. For contributions in the vein, see Glazebrook [3] and Glazebrook and Gittins [4]. All the results known in this area to date require that in some sense the future prospects of the jobs improve indefinitely as they are processed. For example, Glazebrook and Gittins prove that when the function

$$E(a^{P_i - x} | P_i \geqslant x + 1) \qquad (3)$$

is nondecreasing in x for each $i \in J$ (this happens if P_i has a nondecreasing hazard rate) there is an optimal strategy for economic criterion (1) given by a fixed permutation of J. However, in many contexts, for example research planning (see Nash [7]), it is rather more realistic to expect that the future prospects of jobs, after an initial (perhaps lengthy) period of improvement, will begin to deteriorate. It is with this in mind that in Section 2 we demonstrate that the above result of Glazebrook and Gittins may be generalized in a way which does not put monotonicity requirements on the function in (3). Some extensions of this result are discussed in Section 3. In Section 4 we demonstrate how the results of Section 2 may be utilized to give an indication of how well an optimal permutation performs relative to an optimal strategy in a wide range of problems of practical interest. We conclude in Section 5 with a simple example involving five jobs.

2. THE MAIN RESULT

We shall consider the problem of finding optimal strategies (in the set of all feasible strategies) for the pair (J, R) when the economic criterion (1) applies. We shall demonstrate that there exists an optimal strategy which is deterministic, stationary, Markov and nonpreemptive (DSMNP), that is which is given by a fixed permutation specifying in which order the jobs are to be done, when the following conditions hold:

CONDITION 1. $m(i, x) > m(i, 0)$, $x \in Z^+$, $i \in J$;

CONDITION 2: $\lim_{x \to \infty} m(i, x)$ exists and is strictly greater than $m(i, 0)$, $i \in J$; the function $m(i, .)$ being defined as follows:

$$p(P_i \geqslant x + 1) > 0 \Rightarrow m(i, x) = E(a^{P_i - x} | P_i \geqslant x + 1);$$

$$p(P_i \geqslant x + 1) = 0 \Rightarrow m(i, x) = 1.$$

Conditions 1 and 2 are more general than those given by Glazebrook and Gittins [4]. Condition 1 states (loosely) that a task is always brought nearer completion by being processed for an arbitrary length of time. As will be demonstrated in Section 4, the results of this section have implications beyond problems in which conditions 1 and 2 are satisfied.

Before proceeding to the proof of our main result, note first that it is a consequence of an important result in the theory of Markov Decision Processes (see, for example, Ross [8]) that there exists an optimal strategy for our problem which is deterministic, stationary and Markov and so we may restrict our analysis to such strategies.

We require some terminology and notation. By the state of an incomplete job we mean the amount of processing it has received. If job i has been completed its state is denoted $*$. We denote by $C(x_1, x_2, \ldots, x_k) \equiv C(\underline{x})$, the total expected cost incurred by all the jobs in a system identical to the one under study except that job j is in general state x_j initially instead of necessarily being in state 0, $j \in J$, the assumption being that an optimal strategy is adopted. $\bar{C}(\underline{x})$ is similarly defined, the assumption now being that an optimal DSMNP strategy is adopted. I denotes the subset of jobs in J which have no predecessors according to R, i.e.,

$$I = \{i;\ i \in J \text{ and } (j, i) \notin R \text{ for any } j \in J\}.$$

Both $C(\underline{x})$ and $\bar{C}(\underline{x})$ may be characterized as the solutions to appropriately formulated dynamic programming optimality equations:

$$C(\underline{x}) = \min_{i \in I} [ap(P_i = x_i + 1 | P_i > x_i)\{- K(i) + C(x_1, \ldots, x_{i-1}, *, x_{i+1}, \ldots, x_k)\}$$

$$+ ap(P_i > x_i + 1 | P_i > x_i) C(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_k)],$$

and

$$\bar{C}(\underline{x}) = \min_{i \in I} [- K(i) m(i, x_i) + m(i, x_i) C(x_1, \ldots, x_{i-1}, *, x_{i+1}, \ldots, x_k)].$$

The following lemma is the key to establishing our main result.

LEMMA 1:

$$C(\underline{x}) \geqslant \left[\prod_{i=1}^{k} m(i, x_i) \{m(i, 0)\}^{-1} \right] \bar{C}(\underline{0})$$

for all states $\underline{x} \in (Z^+ \cup \{0\})^k$ such that

(i) $i \in I \Rightarrow x_i \geqslant 0$

(ii) $i \notin I \Rightarrow x_i = 0$

(iii) $m(i, x_i) < 1,\ i \in J.$

PROOF: The proof is by means of an induction on K. The lemma clearly holds when $K = 1$ since $C(x_1) = - K(1) m(1, x_1)$ and $\bar{C}(0) = - K(1) m(1, 0)$. We assume that the lemma holds for an arbitrary problem with $K = r - 1$ and demonstrate its validity when $K = r$.

Hence, we consider a problem with r jobs where the position at time 0 is that no jobs have been completed and job i has been processed for x_i units of time where $m(i, x_i) < 1$ and $x_i > 0 \Rightarrow i \in I$. Let S be an optimal strategy for this problem.

AD-A100 442   OFFICE OF NAVAL RESEARCH ARLINGTON VA   F/G 12/2
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 28. NUMBER 2,(U)
JUN 81 S M SELIG

UNCLASSIFIED   NL

2 OF 2
AD A
100442

Suppose that at time 0, S chooses to process job 1 $(\in I)$, then

$$
\begin{aligned}
(4)\qquad C(\underset{\sim}{x}) &= ap(P_1 = x_1 + 1 | P_1 > x_1)\{-K(1) + C(*_1, x_2, \ldots, x_r)\} \\
&\quad + ap(P_1 > x_1 + 1 | P_1 > x_1) C(x_1 + 1, x_2, \ldots, x_r) \\
&= ap(P_1 = x_1 + 1 | P_1 > x_1)\{m(1, x_1)\}^{-1}\{-K(1)m(1, x_1) \\
&\qquad + m(1, x_1)C(*_1, x_2, \ldots, x_r)\} \\
&\quad + ap(P_1 > x_1 + 1 | P_1 > x_1) m(1, x_1 + 1)\{m(1, x_1)\}^{-1} \\
&\qquad [C(x_1 + 1, x_2, \ldots, x_r) m(1, x_1)\{m(1, x_1 + 1)\}]^{-1}.
\end{aligned}
$$

Now by our inductive hypothesis

$$
\begin{aligned}
& -K(1)m(1, x_1) + m(1, x_1)C(*_1, x_2, \ldots, x_r) \\
& \geqslant -K(1)m(1, x_1) + m(1, x_1)\left[\prod_{i=2}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(*_1, 0, \ldots, 0) \\
(5)\qquad & \geqslant \left[\prod_{i=1}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right][-K(1)m(1, 0) + m(1, 0)\bar{C}(*_1, 0, \ldots, 0)] \\
(6)\qquad & \geqslant \left[\prod_{i=1}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(\underset{\sim}{0}),
\end{aligned}
$$

(5) following from Condition 1 and (6) from the fact that the expression in the square brackets in (5) is the expected total cost incurred by the DSMNP strategy which first processes job 1 $(\in I)$ to completion and which after that first completion, processes according to an optimal permutation for the jobs $J - \{1\}$.

We also have that

$$
\begin{aligned}
(7)\qquad & ap(P_1 = x_1 + 1 | P_1 > x_1)\{m(1, x_1)\}^{-1} \\
& \quad + ap(P_1 > x_1 + 1 | P_1 > x_1) m(1, x_1 + 1)\{m(1, x_1)\}^{-1} = 1
\end{aligned}
$$

and so, from (4), (6) and (7), in order to establish that

$$
(8)\qquad C(\underset{\sim}{x}) \geqslant \left[\prod_{i=1}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(\underset{\sim}{0}),
$$

it is sufficient to demonstrate that we must either have

$$
(9)\qquad p(P_1 > x_1 + 1 | P_1 > x_1) = 0
$$

or that

$$
C(x_1 + 1, x_2, \ldots, x_r) m(1, x_1)\{m(1, x_1 + 1)\}^{-1} \geqslant \left[\prod_{i=1}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(0).
$$

That is, that

$$
(10)\qquad C(x_1 + 1, x_2, \ldots, x_r) \geqslant m(1, x_1 + 1)\{m(1, 0)\}^{-1}\left[\prod_{i=2}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(0).
$$

To summarize, in order to establish the desired inequality (8) for state $(x_1, x_2, \ldots, x_r)$ it is sufficient to establish the corresponding inequality for state $(x_1 + 1, x_2, \ldots, x_r)$ this latter state being the result at time $t = 1$ of applying optimal strategy S to the process at time $t = 0$, given that no job completion occurs before time $t = 1$. Should a job completion occur (which will be job 1) before $t = 1$ with probability 1 then inequality (8) is satisfied.

We define N^* as follows:

$$N^* = \inf_{N \in Z^+} \{N;\ \text{with probability one the application of optimal strategy } S \text{ during } [0, N)$$
$$\text{results in at least one job completion, the initial state being } \underset{\sim}{x}\}.$$

We further define $\underset{\sim}{x}(N)$, $0 \leqslant N < N^*$, to be the state resulting at time $t = N$ from the application of optimal strategy S to the process from time $t = 0$ when the initial state is $\underset{\sim}{x}$, given that no job completion occurs during $[0, N)$. For example, if $N^* \geqslant 1$ then $\underset{\sim}{x}(1) = (x_1 + 1, x_2, \ldots, x_r)$.

By repetition of the argument in the paragraph following (10) it is clear that in order to establish (8) it is sufficient to demonstrate that we must have either (i) or (ii).

(i) $N^* < \infty$.

In this case, it is not difficult to show we must have

$$p\{P_j > x_j(N^* - 1) + 1 | P_j > x_j(N^* - 1)\} = 0$$

where j is the job chosen by S for processing during $[N^* - 1, N^*)$ assuming that no job has been completed prior to $N^* - 1$. Hence, referring back to (9), in the case $N^* < \infty$, (8) is established and the induction goes through.

(ii) $N^* = \infty$ and

$$(11) \qquad C\{\underset{\sim}{x}(N)\} \geqslant \left[\prod_{i=1}^{r} m(i, x_i(N))\{m(i, 0\}^{-1}\right] \bar{C}(\underset{\sim}{0})$$

for some $N \in Z^+ \cup \{0\}$.

Hence, we now assume that $N^* = \infty$ (that is, that we cannot be certain of a job completion under S in any particular finite time interval) and consider two cases.

CASE 1: $\underset{\sim}{x}(N)$ has a single positive component ($x_l(N)$, say) for all $N \in Z^+ \cup \{0\}$. When this is so we have that

$$(12) \qquad C(\underset{\sim}{x}) = -K(l)m(l, x_l) + m(l, x_l)C(x_1, \ldots, x_{l-1}, *_l, x_{l+1}, \ldots, x_r)$$
$$= -K(l)m(l, x_l) + m(l, x_l)C(0, \ldots, *_l, \ldots 0)$$
$$(13) \qquad \geqslant m(l, x_l)\{m(l, 0)\}^{-1}\bar{C}(\underset{\sim}{0})$$
$$= \left[\prod_{i=1}^{r} m(i, x_i)\{m(i, 0)\}^{-1}\right]\bar{C}(\underset{\sim}{0}),$$

as required, (12) and (13) following since $x_i = 0$, $i \neq l$.

CASE 2: $\underset{\sim}{x}(N)$ has at least two positive components for all $N \geqslant \hat{N}$, say. When this is the case it follows from Conditions 1 and 2 that

$$(14)\qquad \lim_{N\to\infty} [-K(i)m(i,\ x_i(N))]$$
$$> \lim_{N\to\infty} \left(-K(i)m(i,\ 0)\left[\prod_{j=1}^{r} m(j,\ x_j(N))\{m(j,\ 0)\}^{-1}\right]\right),\ 1 \leqslant i \leqslant r,$$

and from the inductive hypothesis that

$$(15)\qquad \lim_{N\to\infty} \inf\ [m(i,\ x_i(N))C\{x_1(N),\ \ldots,\ *_i,\ \ldots,\ x_r(N)\}]$$
$$\geqslant \lim_{N\to\infty} (m(i,\ 0)\left[\prod_{j=1}^{r} m(j,\ x_j(N))\{m(j,\ 0)\}^{-1}\right]\bar{C}(0,\ \ldots,\ *_i,\ \ldots,\ 0)),\ i \in I.$$

It follows from (14) and (15) that

$$(16)\qquad C^i \underline{\Delta} \lim_{N\to\infty} \inf\ [-K(i)m(i,\ x_i(N))$$
$$+\ m(i,\ x_i(N))C\{x_1(N),\ \ldots,\ *_i,\ \ldots,\ x_r(N)\}]$$
$$> \lim_{N\to\infty} \left(\left[\prod_{j=1}^{r} m(j,\ x_j(N))\{m(j,\ 0)\}^{-1}\right]\{-K(i)m(i,\ 0)\right.$$
$$\left.+\ m(i,\ 0)\bar{C}(0,\ \ldots,\ *_i,\ \ldots,\ 0)\}\right)$$
$$\geqslant \lim_{N\to\infty} \left(\left[\prod_{j=1}^{r} m(j,\ x_j(N))\{m(j,\ 0)\}^{-1}\right]\bar{C}(0)\right)\ i \in I.$$

Let $\bar{N} \in Z^+$ and $\epsilon > 0$ be such that for $N \geqslant \bar{N}$

$$(17)\qquad -K(i)m(i,\ x_i(N)) + m(i,\ x_i(N))C\{x_1(N),\ \ldots,\ *_i,\ \ldots,\ x_r(N)\}$$
$$\geqslant C^i - \epsilon,\ i \in I,$$

and

$$(18)\qquad m(i,\ x_i(N+s))\{m(i,\ x_i(N))\}^{-1} \geqslant (1+\epsilon)^{-1},\ s \in Z^+ \cup \{0\},\ i \in I.$$

We shall now demonstrate that for $N \geqslant \bar{N}$

$$(19)\qquad C\{\underline{x}(N)\} \geqslant \left(\min_{i \in I} C^i - \epsilon\right)(1+\epsilon)^r$$

and, hence, that

$$(20)\qquad \lim_{N\to\infty} \inf\ [C\{\underline{x}(N)\}] \geqslant \min_{i \in I} C^i.$$

Having established (20) it will then follow from (16) that

$$\lim_{N\to\infty} \inf\ [C\{\underline{x}(N)\}] > \lim_{N\to\infty} \left(\left[\prod_{i=1}^{r} m(i,\ x_i(N))\{m(i,\ 0)\}^{-1}\right]\bar{C}(0)\right)$$

from which follows the existence of an $N \in Z^+ \cup \{0\}$ for which (11) holds. This established, the induction will go tnrough and the lemma follows.

We now proceed to demonstrate (19). We consider a problem where at $t = 0$ no job is complete and that tne state of the process is $\underline{x}(N)$, $N \geqslant \bar{N}$. Suppose that optimal strategy S indicates that at time $t = 0$ $(= t_0)$ task j_1 should be processed until time $t = t_1 (> t_0)$ or until j_1 is completed, whichever occurs sooner. At time $t = t_1$, if j_1 has not been completed, optimal strategy S indicates that job j_2 $(\neq j_1)$ should be processed until time $t = t_2$ $(> t_1)$ or until j_2 is completed, whichever occurs sooner, and so on. Under the assumption that no job is completed before time $t = t_{n-1}$, S indicates that job j_n ($\neq j_{n-1}$) should be processed until time $t = t_n$ ($>\_t_{n-1}$) or until j_n is completed, whichever occurs sooner, $1 \leqslant n < \infty$. It is clear that for $N \geqslant \bar{N}$

$$\begin{aligned} C\{\underset{\sim}{x}(N)\} = \sum_{n=0}^{\infty} a^{t_{n-1}} \prod_{i=1}^{r} p\{P_i > x_i(N + t_{n-1}) \mid P_i > x_i(N)\} \\ \times \sum_{s=1}^{t_n - t_{n-1}} a^s p\{P_{j_n} = x_{j_n}(N + t_{n-1} + s) \mid P_{j_n} > x_{j_n}(N + t_{n-1})\}[- K(j_n) \\ + C\{x_1(N + t_{n-1}), \ldots, *_{j_n}, \ldots, x_r(N + t_{n-1})\}] \end{aligned}$$

which, by (18), is

$$\begin{aligned} \geqslant (1+\epsilon)^r \sum_{n=0}^{\infty} \Bigg\{\Bigg[\prod_{i=1}^{r} m(i, x_i(N + t_{n-1})) \\ \{m(i, x_i(N))\}^{-1}\Bigg] p\{P_i > x_i(N + t_{n-1}) \mid P_i > x_i(N)\}] a^{t_{n-1}} \\ \times \sum_{s=1}^{t_n - t_{n-1}} a^s p\{P_{j_n} = x_{j_n}(N + t_{n-1} + s) \mid P_{j_n} > x_{j_n}(N + t_{n-1})\} \\ \{m(j_n, x_{j_n}(N + t_{n-1}))\}^{-1} \\ \times [- K(j_n) m(j_n, x_{j_n}(N + t_{n-1})) \\ + m(j_n, x_{j_n}(N + t_{n-1})) C\{x_1(N + t_{n-1}), \ldots, *_{j_n}, \ldots, x_r(N + t_{n-1})\}]\Bigg\}; \end{aligned}$$

which, by (17), is

$$\begin{aligned} (21) \qquad \geqslant (1+\epsilon)^r \left\{\min_{i \in I} C^i - \epsilon\right\} \sum_{n=0}^{\infty} \Bigg\{\Bigg[\prod_{i=1}^{r} m(i, x_i(N + t_{n-1})\Bigg] \\ \{m(i, x_i(N))\}^{-1} p\{P_i > x_i(N + t_{n-1}) \mid P_i > x_i(N)\}\Bigg] \\ \times a^{t_{n-1}} \sum_{s=1}^{t_n - t_{n-1}} a^s p\{P_{j_n} = x_{j_n}(N + t_{n-1} + s) \mid P_{j_n} > x_{j_n}(N + t_{n-1})\} \\ \{m(j_n, x_{j_n}(N + t_{n-1}))\}^{-1}\Bigg\} \\ = (1+\epsilon)^r \left\{\min_{i \in I} C^i - \epsilon\right\}, \end{aligned}$$

since the infinite sum in (21) can be shown to be one (the proof is based on (7)). We have thus established (19) and hence the induction goes through and the lemma follows.

THEOREM 1: There is a DSMNP strategy which is optimal.

PROOF: We may take $x_i = 0$, $i \in J$, in Lemma 1 in which case we obtain that

$$C(\underset{\sim}{0}) \geqslant \bar{C}(\underset{\sim}{0}).$$

Theorem 1 follows immediately.

3. EXTENSIONS AND COMMENTS

(3.1) Weak Conditions

Theorem 1 continues to hold when the strict inequalities in Conditions 1 and 2 are replaced by weak ones as follows:

CONDITION 1′: $m(i, x) \geqslant m(i, 0)$, $x \in Z^+$, $i \in J$;

CONDITION 2′: $\lim_{x \to \infty} m(i, x)$ exists, $i \in J$

The proof combines the results in Section 2 with a truncation argument of a kind which will be used in Section 4.

(3.2) Linear Costs

It is frequently the case (see, for example, *Glazebrook* [3]) that results for problems with linear costs may be deduced from equivalent results for problems with discounted costs by means of arguments which involve allowing the discount rate to tend to one. Suppose we consider the problem outlined in Section 1 with costs given by (2). It may be deduced from the results in the previous section (together with paragraph (3.1)) that under the conditions:

CONDITION 1″: $n(i, x) \leqslant n(i, 0)$, $x \in Z^+$, $i \in J$;

CONDITION 2″: $\lim_{x \to \infty} n(i, x)$ exists, $i \in J$, where

$$p(P_i \geqslant x+1) > 0 \Rightarrow n(i, x) = E(P_i - x | P_i \geqslant x+1)$$

$$p(P_i \geqslant x+1) > 0 \Rightarrow n(i, x) = 0$$

there exists an optimal strategy which is DSMNP. This is a generalization of a result due to Glazebrook and Gittins [4].

(3.3) Continuous Time Analogues

For simplicity our discussion is restricted to discrete time problems. Continuous time analogues of the main results may be obtained by means of delicate limiting arguments, considering optimal strategies for appropriately chosen sequences of discrete time problems, allowing the discrete time quantum to tend to zero.

(3.4) Algorithm Selection

Once we have established that a problem has an optimal strategy which is DSMNP, the question arises of which permutation (or permutations) determines this optimal strategy. An algorithm which generates the appropriate permutation for discounted costs (1) may be found in Glazebrook and Gittins [4]; an algorithm for the linear costs case (2) is to be found in Sidney [10].

4. THE EVALUATION OF NONPREEMPTIVE STRATEGIES

Conditions 1(1', 1") and 2(2', 2"), though they take us much further than the monotonicity requirements of previous work, do limit the range of direct application of the material in Section 2. The main limitation is in the insistence that jobs should always be at least as promising (i.e., always have at least as low an expected remaining cost) as they are initially. However, it turns out that the results of Section 2, though limited in this way in their direct application, help us in the important task of evaluating how well an optimal DSMNP strategy performs relative to an optimal strategy in a large class of problems of practical interest.

As was implied in the introduction, even if a stochastic job cannot be assumed *always* to be at least as promising as it is initially then in many practical contexts such an assumption can at least be valid for some initial phase of the job's development. For some examples of this, see Nash [7] whose interest is in modeling research projects and Singh and Billinton [11] who commend the lognormal distribution as a good model for repair times. Such considerations motivate the following definitions:

DEFINITION 1: Job i is said to be *initially improving for the discounted costs problem* if $m(i, 1) \geqslant m(i, 0)$ and if $\lim\limits_{x \to \infty} m(i, x)$ *exists.*

DEFINITION 2: Job i is said to be *initially improving for the linear costs problem* if $n(i, 1) \leqslant n(i, 0)$ and if $\lim\limits_{x \to \infty} n(i, x)$ exists.

(4.1) Discounted costs

Throughout this subsection we shall assume that all jobs in J are initially improving for the discounted costs problem. We shall also assume economic criterion (1).

We define

$$t_i = \sup_{t \in Z^+} \{t;\ m(i, x) \geqslant m(i, 0), 0 \leqslant x \leqslant t\},\ i \in J. \tag{22}$$

We further define the *random* variable P_i^* to be the processing time P_i truncated at $T_i + 1$. Corresponding to P_i^* is the function $m^*(i, .)$. The following lemma is easy to establish.

LEMMA 2:

(i) $m^*(i, x) \geqslant m^*(i, 0),\ x \in Z^+,\ i \in J.$

(ii) $\lim\limits_{x \to \infty} m^*(i, x)$ exists, $i \in J.$

Hence, the truncated processing time P_i^* satisfies Conditions 1' and 2' of paragraph (3.1). Now, the main idea of this section is as follows: suppose that for each job $i \in J$, T_i is large (which in many practical problems it will be); then the total expected cost incurred by an optimal strategy will be close to the total expected cost incurred by an optimal strategy for the equivalent problem with the processing time $\{P_1, P_2, \ldots, P_K\}$ replaced by the truncated processing times $\{P_1^*, P_2^*, \ldots, P_K^*\}$. However, from Theorem 1 and Lemma 2 this latter problem has an optimal strategy which is DSMNP. These considerations lead us to expect an optimal DSMNP strategy to perform well relative to an optimal strategy. Theorem 2 aims to quantify these ideas.

THEOREM 2:

$$\{C(\underset{\sim}{0}) - \bar{C}(\underset{\sim}{0})\}\{C(\underset{\sim}{0})\}^{-1} \leqslant \left[\prod_{i=1}^{K} m^*(i,\ 0)\{m(i,\ 0)\}^{-1}\right] - 1.$$

PROOF: Let an optimal DSMNP strategy for the problem with processing times $\{P_1,\ P_2,\ \ldots,\ P_K\}$ replaced by truncated times $\{P_1^*,\ P_2^*,\ \ldots,\ P_K^*\}$ be given by the permutation $\{\alpha(1),\ \alpha(2),\ \ldots,\ \alpha(K)\}$. By Theorem 1 and Lemma 2 this strategy is optimal for that problem in the class of all feasible strategies. Let $C^*(\underset{\sim}{0})$ be the expected total cost incurred by the application of this strategy to the problem with the truncated processing times and let $\hat{C}(\underset{\sim}{0})$ be the expected total cost incurred by the application of this same permutation to the original problem with nontruncated processing times. It is clear that

$$0 > \hat{C}(\underset{\sim}{0}) \geqslant \bar{C}(\underset{\sim}{0}) \geqslant C(\underset{\sim}{0}) \geqslant C^*(\underset{\sim}{0}).$$

Hence,

$$\begin{aligned}\{C(\underset{\sim}{0}) - \bar{C}(\underset{\sim}{0})\}\{C(\underset{\sim}{0})\}^{-1} &\leqslant \{C^*(\underset{\sim}{0}) - \hat{C}(\underset{\sim}{0})\}\{\hat{C}(\underset{\sim}{0})\}^{-1}\\ &= \sum_{i=1}^{K} - K\{\alpha(i)\}\left\{\prod_{j=1}^{i} m^*(\alpha(j),\ 0) - \prod_{j=1}^{i} m(\alpha(j),\ 0)\right\}\\ &\qquad\left[\sum_{i=1}^{K} - K\{\alpha(i)\}\prod_{j=1}^{i} m(\alpha(j),\ 0)\right]^{-1}\\ &\leqslant \left\{\prod_{i=1}^{K} m^*(\alpha(i),\ 0) - \prod_{i=1}^{K} m(\alpha(i),\ 0)\right\}\left\{\prod_{i=1}^{K} m(\alpha(i),\ 0)\right\}^{-1}\\ &= \left[\prod_{i=1}^{K} m^*(i,\ 0)\{m(i\ \ 0)\}^{-1}\right] - 1,\end{aligned}$$

as required.

(4.2) Linear costs

Throughout this subsection we shall assume that all jobs in J are initially improving for the linear costs problem. Costs $C(\underset{\sim}{0})$ and $\bar{C}(\underset{\sim}{0})$ are as in (4.1) except that now they refer to economic criterion (2).

We define as before

$$S_i = \sup_{t \in Z^+} \{t;\ n(i,\ x) \leqslant n(i,\ 0),\ 0 \leqslant x \leqslant t\} \tag{23}$$

and thus obtain function $n^*(i,\ .)$ as in (4.1). This function is found to satisfy Conditions 1″ and 2″ and so we have Theorem 3.

THEOREM 3:

$$\{\bar{C}(\underset{\sim}{0}) - C(\underset{\sim}{0})\}\{C(\underset{\sim}{0})\}^{-1} \leqslant \max_{1 \leqslant i \leqslant K} [\{n(i,\ 0) - n^*(i,\ 0)\}\{n^*(i,\ 0)\}^{-1}].$$

PROOF: The proof is similar to Theorem 2.

We deduce from Theorems 2 and 3 that when dealing with collections of initially improving jobs whose associated values of T_i and S_i are large we lose little by restricting our attention to DSMNP strategies. Note too, that in any given problem it may be that we can truncate at times considerably larger than $T_i + 1$ or $S_i + 1$ and still have functions $m^*(i, .)$ or $n^*(i, .)$ satisfying the appropriate conditions. When this is the case it may be possible to improve the bounds given in *Theorems 2 and 3.*

Note further that Theorems 2 and 3 also hold in continuous time. The modifications required are that in the definitions of T_i and S_i in (22) and (23) respectively the suprema should be taken over R^+, the nonnegative real numbers, and that to obtain P_i^* in both cases, truncations are taken at T_i and S_i respectively. We also need to modify Definitions 1 and 2 in the obvious way.

5. EXAMPLE

For simplicity, we consider an example in continuous time with linear costs as in (2). There are five jobs and so $J = \{1,2,3,4,5\}$ with predence relation $R = \{(1,2), (1,5), (2,3), (5,3)\}$. It is not difficult to see that there are ten feasible DSMNP strategies for J. The distribution of P_i is summarized by its hazard rate $\lambda_i(.)$ which is assumed to have the form

$$\lambda_i(x) = \begin{cases} \lambda_{1i}, & 0 \leqslant x < T_{1i}, \\ \lambda_{2i}, & T_{1i} \leqslant x < T_{1i} + T_{2i}, \qquad i = 1,2,3,4,5. \\ \lambda_{3i}, & T_{1i} + T_{2i} \leqslant x, \end{cases} \tag{24}$$

The important details for the five jobs are summarized in Table 1. It is easy to show, by application of the algorithm due to Sidney [10] that the optimal permutation is (4,1,5,2,3) with associated expected cost $\underset{\sim}{C}(\underline{0}) = 31.089$.

TABLE 1

Job (i)	$K(i)$	λ_{1i}	λ_{2i}	λ_{3i}	T_{1i}	T_{2i}	$n(i, 0)$
1	1	1	3	2	1	1	0.758
2	2	1	3	1.5	1	2	0.817
3	3	2	5	2.5	2	4	0.495
4	4	2	4	3	2	1	0.495
5	5	1	2	1	3	3	0.975

It is also not difficult to demonstrate that, with processing time distributions given according to (24) that

$$\lambda_{2i} \geqslant \lambda_{1i} \quad \text{and} \quad \lambda_{3i} \geqslant \lambda_i^* \tag{25}$$

where

$$(\lambda_i^*)^{-1} = (\lambda_{1i})^{-1}\{1 - \exp(-\lambda_{1i}T_{1i})\} + (\lambda_{2i})^{-1}\{\exp(-\lambda_{1i}T_{1i}) - \exp(-\lambda_{1i}T_{1i} - \lambda_{2i}T_{2i})\}$$
$$\times \{1 - \exp(-\lambda_{1i}T_{1i} - \lambda_{2i}T_{2i})\}^{-1}$$

are sufficient to ensure that

$$n(i, x) \leqslant n(i, 0), \; x \in R^+,$$

and the existence of

$$\lim_{x \to \infty} n(i, x).$$

Jobs 1, 2, 3 and 4 all satisfy (25) but job 5 does not. Indeed,

$$n(5,\ x) = 1 > n(5,\ 0),\ x > 6.$$

However, job 5 is initially improving in the sense that the (right-hand) derivative of $n(5,\ x)$ at $x = 0$ is negative, and so the theory of Section 4 applies. In fact, the value S_5 can be shown to be 5.975 and the continuous-time version of Theorem 3 applied to this case yields

$$\{\bar{C}(\underset{\sim}{0}) - C(\underset{\sim}{0})\}\{C(\underset{\sim}{0})\}^{-1} \leqslant \{n(5,\ 0) - n^*(5,\ 0)\}\{n^*(5,\ 0)\}^{-1} = 1.30 \times 10^{-4},$$

whereupon we obtain, that

$$31.085 \leqslant C(\underset{\sim}{0}) \leqslant 31.089.$$

Evidently, then, very little is lost in this case by restricting attention to permutations of J.

REFERENCES

[1] Garey, M.R., "Optimal Task Sequencing with Precedence Constraints," Discrete Mathematics, *4*, 37-56 (1973).

[2] Glazebrook, K.D., "Stochastic Scheduling with Order Constraints," International Journal of Systems Science, *7*, 657-666 (1976).

[3] Glazebrook, K.D. "On Stochastic Scheduling with Precedence Relations and Switching Costs," Journal of Applied Probability, *17*, 1016-1024 (1980).

[4] Glazebrook, K.D. and J.C. Gittins, "On Single-Machine Scheduling with Precedence Relations and Linear on Discounted Costs," Operations Research, *29*, (1981, to appear).

[5] Meilijson, I. and G. Weiss, "Multiple Feedback at a Single Server Station," Stochastic Processes and their Applications, *5*, 195-205 (1977).

[6] Monma, C.L. and J.B. Sidney, "Sequencing with Series-Parallel Precedence Constraints," (submitted for publication).

[7] Nash, P., "Optimal Allocation of Resources between Research Projects," Ph.D. Thesis, Cambridge University, Cambridge, England (1973).

[8] Ross, S.M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco, Calif., 1970).

[9] Rothkopf, M.E., "Scheduling Independent Tasks on Parallel Processors," Management Science, *12*, 437-447 (1966).

[10] Sidney, J.B., "Decomposition Algorithms for Single Machine Sequencing with Precedence Relations and Deferral Costs," Operations Research, *23*, 283-298 (1975).

[11] Singh, C. and R. Billinton, *System Reliability-Modelling and Evaluation*, (Hutchinson & Co., London, England, 1977).

POSTOPTIMALITY ANALYSIS IN NONLINEAR INTEGER PROGRAMMING: THE RIGHT-HAND SIDE CASE

Mary W. Cooper

Department of Operations Research and
Engineering Management
Southern Methodist University
Dallas, Texas

ABSTRACT

An algorithm is presented to gain postoptimality data about the family of nonlinear pure integer programming problems in which the objective function and constraints remain the same except for changes in the right-hand side of the constraints. It is possible to solve such families of problems simultaneously to give a global optimum for each problem in the family, with additional problems solved in under 2 CPU seconds. This represents a small fraction of the time necessary to solve each problem individually.

1. INTRODUCTION

Recently efforts have been made to extend the ideas of postoptimal analysis and parametric analysis which are widely used in linear programming to 0-1 integer programming and general integer programming. A review of these efforts is given by Geoffrion and Nauss [4]. They cite work on the 0-1 problem by G. Roodman [13], and an extension of that work by Piper and Zoltners [11]. Roodman [12] and Marsten and Morin [8] have looked at the same topic using branch and bound. These and other authors are cited in [4]. Bailey and Gillett [1] have recently used cutting planes in parametric integer programming. The present paper differs from these efforts in considering postoptimal right-hand side analysis for a different problem: the pure integer nonlinear programming problem with separable objective function and constraints. Our purpose is to modify an algorithm which has been previously described [3] so that it simultaneously finds optimal solutions for a family of problems of the type described above which differ only in the right-hand side vector of the constraints. (This family is analagous to Geoffrion and Nauss' family P_u in their discussion of postoptimality analysis for the linear integer case).

2. APPLICATIONS

One of the most general formulations to which this algorithm applies is the separable nonlinear knapsack problem. It has numerous application areas in allocation of resources, cutting stock problems and capital budgeting [7], [9], [10], [5], [6]. In addition it has applications for solving subproblems in many integer programming algorithms [14], [2], [15]. The importance of the work in this report which gives postoptimality data for this problem can be argued in a way analagous to the case for linear programming. Additional information about the value of changes in resources, is usually worth a minor amount of additional computation. Often right-hand side values represent estimates, and information about the effect of right-hand side changes on the optimal solution represents a crude determination of the effect of estimating a variable by its expected value.

3. THE PROBLEM AND METHOD

Let us first characterize the problems we solve, and second, briefly review the elements of the algorithm to be modified. After these sections, the algorithm is extended to solve the family of problems which differ only in the right-hand side vector.

Let us use the following notation to formulate the problem (P).

$$\text{(1)} \qquad \text{Maximize } z = \sum_{j=1}^{n} f_j(x_j) \text{ subject to}$$

$$\text{(2)} \qquad \sum_{j=1}^{n} h_{ij}(x_j) \leqslant b_i, \; i = 1, 2, \ldots, m, \text{ and } x_j \in I_p \text{ for } j = 1, \ldots, n.$$

Additional restrictions on the functions are

(1) $f_j : I_p \to R_p$, $j = 1, \ldots, n$, and they satisfy a sufficient condition for dynamic programming.

(2) $h_{ij} : I_p \to R_p$, $j = 1, \ldots, n$, and $i = 1, \ldots, m$ and are nondecreasing in x_j.

(3) the region described by the constraints is nonempty, contains at least one integer point, and is bounded.

Our previous algorithm [3] is a top-down enumerative method for solving this problem in which the constraints are used to eliminate infeasible partial solutions and their completions. In this paper we require the additional restriction described above in condition (2), although the paper cited in [3] treats a more general nonseparable form of the constraints. Let us describe the solution process for the pure integer nonlinear separable programming problem given in (1) and (2).

Step 1: Find upper bounds on x_j, $j = 1, \ldots, n$ and z_0 over the constraints in set (2).

Step 2: Solve the following dynamic programming problem:

$$\text{(3)} \qquad \text{Maximize} \qquad Z = \sum_{j=1}^{n} f_j(x_j)$$

$$\text{Subject to} \qquad \sum_{j=1}^{n} f_j(x_j) \leqslant z_0.$$

This *single* dynamic programming problem can be used to identify lattice points on the hypersurface

$$\sum_{j=1}^{n} f_j(x_j) = z_0$$

and on every hypersurface

$$(4) \qquad \sum_{j=1}^{n} f_j(x_j) = z, \; 0 \leqslant z \leqslant z_0.$$

Step 3: We use the dynamic programming solution table to generate *both* a sequence of decreasing values of z which correspond to hypersurface levels containing integer points and also to generate all lattice points on that particular hypersurface. For details of the method, see [3].

Step 4: The constraints (2) of the original problem are used to check for feasibility. The argument is simply, if we look at all hypersurfaces (4) in decreasing order of z, then the first feasible point with respect to the constraints (4) will be optimal.

Actually the feasibility of the solutions is checked at the partial solution stage. For a given z, say z_k, we generate the components of the lattice point in the order x_n^*, x_{n-1}^*, . . . , x_1^*. After x_n^* is generated, the vector corresponding to the remaining resource levels, that is,

$$\bar{b}' = \bar{b} - \bar{a}_n x_n^*$$

is checked for any negative components. If none are found, this partial solution is *still a candidate* for a feasible solution. Otherwise it is eliminated before any other components x_{n-1}^*, x_{n-2}^*, . . . , x_1^* are generated from the dynamic programming tables, since the final solution is infeasible no matter what the remaining components are. Hence, solutions are eliminated from consideration as quickly as possible.

4. ADDITIONAL CALCULATIONS TO DETERMINE OPTIMAL SOLUTIONS FOR CERTAIN MEMBERS OF THE P_θ FAMILY

Let us assume that we want to find optimal solutions to the following problem P_θ

$$\text{Maximize} \qquad z = \sum_{j=1}^{n} f_j(x_j)$$

$$\text{Subject to} \qquad \sum_{j=1}^{n} h_{ij}(x_j) \leqslant b_i + \theta r_i \; i = 1, \; \ldots, \; m.$$

$$x_j \in I_p \text{ for } j = 1, \; \ldots, \; n.$$

$$0 = \theta_0 < \theta_1 < \ldots < \theta_l = 1$$

$$r_i \geqslant 0, \; i = 1, \; \ldots, \; m.$$

Then Step 4 must be changed to include additional tests for feasibility for each of the right-hand side vectors, $\bar{b}_0 = \bar{b}$, $\bar{b}_1 = \bar{b} + \theta_1 \; \bar{r}$, $\bar{b}_2 = \bar{b} + \theta_2\bar{r}$, $\bar{b}_3 = \bar{b} + \theta_3\bar{r}$, . . . , $\bar{b}_l = \bar{b} + \bar{r}$. *Note that* if $0 < \theta_1 < \theta_2 \ldots < 1$, the following relationship between the right-hand side values exists—

$$b_i < b_{i1} = b_{i0} + \theta_1 r_i < b_{i2} = b_{i0} + \theta_2 r_i < \ldots < b_{il} = b_i + r_i$$
$$\text{for } i = 1, \; \ldots, \; m.$$

Let us assume that we are testing the feasibility of a partial solution with constraint i. Then if feasibility is tested for b_{il}, b_{il-1}, . . . , b_1, b_0, if any constraint is violated whose ith constraint has right-hand side value b_{ip}, then for the problems whose right-hand side values are $b_{i,p-1}$, $b_{i,p-2}$, . . . , b_{i1}, the current partial solution will also be infeasible. This is the order of calculation that has been implemented in a computer program. It is also possible to describe an algorithm for solving a set of problems whose right-hand side vectors are not related as those are in the P_u family which decrease in every component. For two problems with arbitrary, and differing right-hand side vectors $\bar{b}_1$ and $\bar{b}_2$, then there may be no method of ordering the $\bar{b}$ vectors so that for every row i, $b_{i1} < b_{i2}$. Hence, a less efficient algorithm could be implemented in which every b_{ij} must be checked, even if an indication of infeasibility is given for a previous $b_{i,j+1}$. The reason is obvious: for arbitrary components no ordering can guarantee that $b_{ij} < b_{i,j+1}$ for every constraint, hence, the ith constraint may not be violated if its right hand side is $b_{i,j}$.

A flow chart of the order of the calculations for implementation of the simultaneous solution of a family of problems differing in the right-hand side is given below. We assume that we have generated an upper bound z_0 on the objective function in some way, and that we are considering a partial solution for some hypersurface with functional value $z_k < z_0$. The assumption is clearly that for all hypersurfaces with intermediate functional values either

(a) they contain no integer points (we do not explicitly consider these), or

(b) they contain no feasible integer points.

At each stage in generating a new component of an integer point from the dynamic programming tables a test for feasibility is made with the new x_i^* and components in the partial solution already obtained. Hence, the flow chart of this part of the algorithm assures that the sequence of functional hypersurface values with integer points has been identified and put in strictly descending order: $z_0 > z_1 > \ldots > z_k > \ldots$. The right-hand side vectors under consideration can be written as

$$\bar{b}_p = \bar{b} + \theta_p \bar{r}, \text{ and } 0 \leqslant \theta_1 < \theta_2 < \theta_p < \theta_l = 1.$$

The program considers the right-hand side vectors in the order $\bar{b}_l, \bar{b}_{l-1}, \ldots, \bar{b}_1, \bar{b}_0$, so that any partial integer solution which is infeasible for $\bar{b}_p$ is also infeasible for all previous right hand-side vectors. The logic is given in the following diagram (Figure 1).

A careful analysis of the program logic will show that many problems of the family P_u can be solved using the solution table from a single dynamic programming problem. We would expect a considerable saving over the time for solving each problem in the family separately for this reason. In addition, the fathoming or discarding of integer points at the partial solution stage can be done for several problems at a time.

5. . ƆMPUTATIONAL DATA

Seven different basic families P_u have been solved on the CDC CYBER 70, Model 72, a moderate speed computer. The results are given in Table 1. m is the number of constraints, n is the number of variables, k is a bound on x_i. Problems are created with randomly generated coefficients. The functions of $f_i(x_i)$ are cubic polynomials, so a problem with 12 variables might have as many as 36 terms in the objective function. Constraints of the form

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i, \; i = 1, \ldots, m$$

are used with the restriction that $a_{ij} \geqslant 0$, $b_i > 0$. For each member of the P_u family, the new right-hand side vector is created by subtracting 5 from each component of $\bar{b}$. A time of .00 indicates that the current optimal solution also solves the next problem in the family which has a smaller value in each component of $\bar{b}$. Note that other schemes of obtaining members of P_u can be easily implemented.

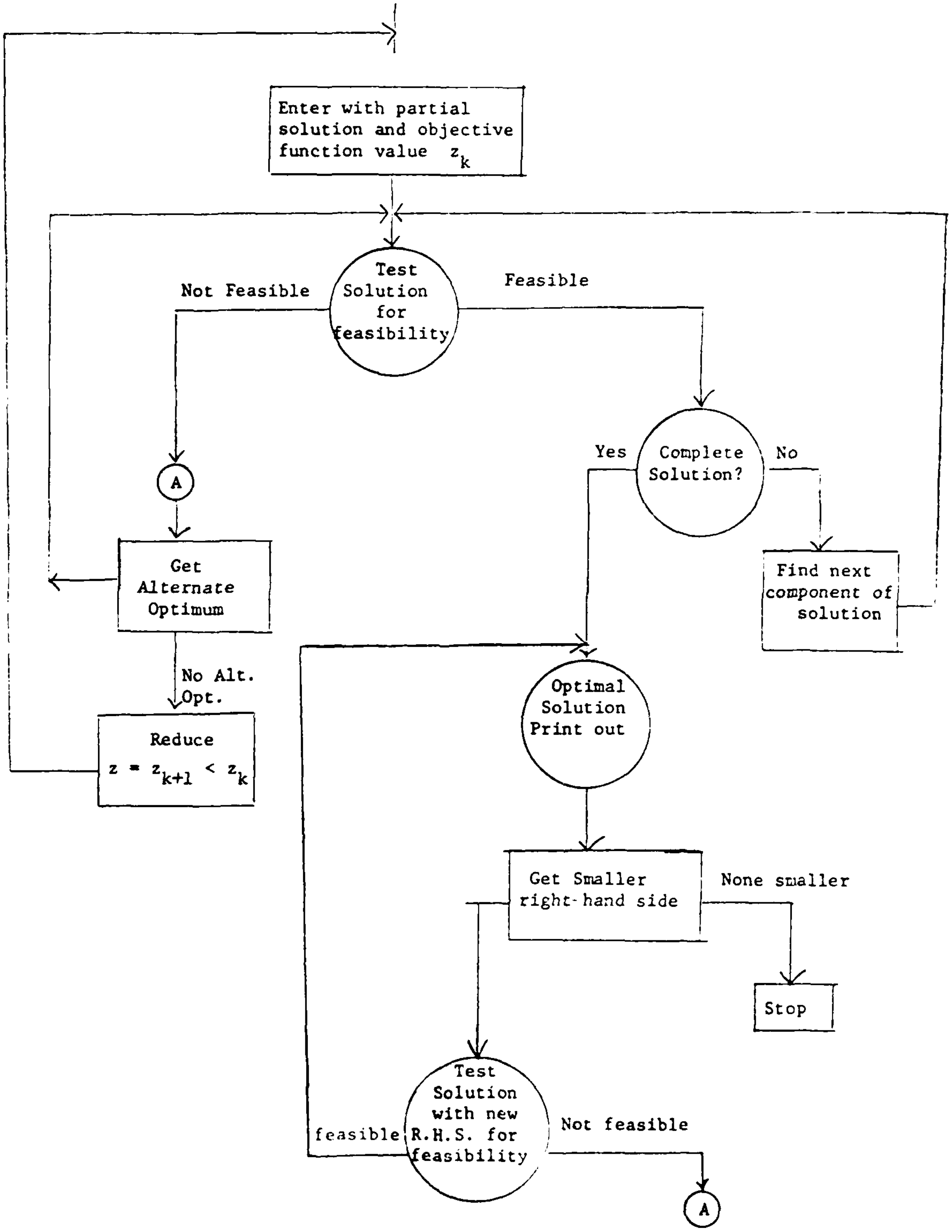

Enter with partial solution and objective function value z_k
Test Solution for feasibility
Not Feasible
Feasible
Complete Solution?
Yes
No
A
Get Alternate Optimum
Find next component of solution
No Alt. Opt.
Reduce $z = z_{k+1} < z_k$
Optimal Solution Print out
Get Smaller right-hand side
None smaller
Stop
Test Solution with new R.H.S. for feasibility
feasible
Not feasible
A

FIGURE 1 Flow chart

TABLE 1

m	n	k	Base Problem ($\bar{b}_1$) (seconds)	Time (CPU Seconds)			
				$\bar{b}_2$	$\bar{b}_3$	$\bar{b}_4$	$\bar{b}_5$
3	12	2	26.43	.00	.55	1.12	.56
4	10	2	13.27	.00	.00	1.38	.44
4	10	2	30.36	.00	.00	.00	1.64
4	15	2	36.26	.00	.00	.00	.00
4	15	2	84.49	1.33	.00	3.87	1.09
4	15	2	32.98	.00	.00	.00	.00
4	15	2	25.70	.00	.00	.00	.66

In each case, although individual problems are between 10-80 CPU seconds, after the base problems are solved, other problems in the same family are solved in under 2 seconds.

6. ACKNOWLEDGMENT

The author would like to acknowledge the valuable comments of T.L. Morin on the topic of this paper.

REFERENCES

[1] Bailey, M.G. and B.E. Gillett, "Parametric Integer Programming Using Cutting Planes," unpublished paper, ORSA/TIMS Meeting, Los Angeles (December 1978).

[2] Bradley, G., "Transformation of Integer Programs to Knapsack Functions," Discrete Mathematics, *1*, 29-45 (1971).

[3] Cooper, M.W., "The Use of Dynamic Programming Methodology for the Solution of a Class of Nonlinear Programming Problems," Naval Research Logistics Quarterly, *27*, No. 1 (1980).

[4] Geoffrion, A.M. and R. Nauss, "Parametric and Postoptimality Analysis in Integer Linear Programming," Management Science, *23*, No. 5 (1977).

[5] Gilmore, P.C. and R.E. Gomory, "Multi-Stage Cutting Stock Problems of Two or More Dimensions," Operations Research, *13*, 94-120 (1965).

[6] Gilmore, P.C. and R.E. Gomory, "The Theory and Computation of Knapsack Functions," Operations Research, *14*, 1045-1074 (1966).

[7] Lorie, J.H. and L.J. Savage, "Three Problems in Rationing Capital," Journal of Business, *28*, 229-239 (1955).

[8] Marsten, R.E. and T.L. Morin, "Parametric Integer Programming: The Right-Hand Side Case," Discrete Mathematics, *1*, 375-390 (1977).

[9] Nemhauser, G.L. and Z. Ullman, "Discrete Dynamic Programming and Captital Allocations," Management Science *15*, 801-810 (1969).

[10] Peterson, C.C., "Computational Experience with Variants of the Belos Algorithm Applied to the Selection of R & D Projects," Management Science, *13*, 736-750 (1967).

[11] Piper, C.J. and A.A. Zoltners, "Some Easy Postoptimality Analysis for Zero-One Programming," Management Science, *22*, No. 7 (1976).

[12] Roodman, G.M., "Postoptimality Analysis in Integer Programming by Implicit Enumeration: The Mixed Integer Case," The Amos Tuck School of Business Administration, Dartmouth College (October 1973).

[13] Roodman, G.M., "Postoptimality Analysis in Zero-One Programming by Implicit Enumeration," Naval Research Logistics Quarterly, *19*, No. 3 (1972).

[14] Salkin, H.H. and C.A. De Kluyver, "The Knapsack Problem: A Survey," Naval Research *Logistics Quarterly*, *22*, 127-144 (1975).

[15] Shapiro, J.V. and H.M. Wagner, "A Finite Renewal Algorithm for the Knapsack and Turnpike Models," Operations Research, *15*, 319-341 (1967).

AN EFFICIENT ALGORITHM FOR THE LOCATION-ALLOCATION PROBLEM WITH RECTANGULAR REGIONS

Ann S. Marucheck

Oklahoma City University
Oklahoma City, Oklahoma

Adel A. Aly

School of Industrial Engineering
University of Oklahoma
Norman, Oklahoma

ABSTRACT

The location-allocation problem for existing facilities uniformly distributed over rectangular regions is treated for the case where the rectilinear norm is used. The new facilities are to be located such that the expected total weighted distance is minimized. Properties of the problem are discussed. A branch and bound algorithm is developed for the exact solution of the problem. Computational results are given for different sized problems.

1. INTRODUCTION

All previous studies of the location-allocation ($L-A$) problem have used the assumption that the location of customers of existing facilities were deterministic points. The multifacility location problem involves the location of one or more new facilities relative to several existing facilities in order to minimize the sum of the weighted distances among the facilities. Previous work [1,2,16] with this problem has shown that in the urban setting, potential location of customers or existing facilities may be more accurately represented as random points uniformly distributed over rectangular regions. Since the $L-A$ problem is a generalized version of the multifacility location problem, the principal of using rectangular regions to represent existing facilities instead of aggregate points would be appropriate in modeling the $L-A$ problem.

A common approach to handling the location problem with rectangular regions is to represent each region by its centroid and to solve the resulting problem as a deterministic model. Although this method is computationally easier, it has been shown [3] that the solutions's proximity to optimality is metric dependent. Location problems with Euclidean distance metric are relatively insensitive to a relaxation of the probabilistic assumptions. In other words, using the centroid approach for probabilistic location problems with Euclidean distance metric yields a near optimal solution. However, the tradeoffs in considering the deterministic (centroid) version of the rectilinear metric location problem are greater [1]. Consequently, in considering probabilistic location formulations using the rectilinear metric it is necessary to develop solution techniques other than the deterministic ones.

Often the solution techniques for the $L-A$ problem involve the use of a facility location algorithm to generate and evaluate allocation schemes. Cooper [6] and Kuenne and Soland [9] both indicate that finding the optimal allocation scheme is the most critical task in solving the $L-A$ problem. Thus, determination of the optimal allocation scheme is only as reliable as the facility location techniques employed.

The purpose of this research effort is to develop and test an exact solution technique for the $L-A$ problem among rectangular regions with a rectilinear metric.

2. FORMULATIONS

The general location-allocation model among rectangular regions is formulated as follows.

$$\text{(P)} \qquad \text{minimize} \sum_{j=1}^{n} \sum_{i=1}^{m} \iint_{R_i} z_{ij} w_i |X_j - R_i|_{l_p} \, \theta(R_i) dR_i$$

$$\text{subject to: } \sum_{j=1}^{n} z_{ij} = 1 \qquad \text{for all } i$$

$$z_{ij} = 0, 1 \qquad \text{for all } i \text{ and } j$$

where: n = number of new facilities
m = number of existing facilities
X_j = (x_j, y_j), coordinate location of new facility j
R_i = existing rectangular region i
$\theta(R_i)$ = bivariate probability density function over R_i
w_i = interaction between region i and the new facility it will be allocated to
z_{ij} = $\begin{cases} 1, \text{ if existing facility } i \text{ is allocated to new facility } j \\ 0, \text{ otherwise} \end{cases}$
l_p = the type of norm used. When $p = 1, 2,$ and ∞, the metric becomes rectilinear, Euclidean, and Chebyshev distances respectively.

The particular problem to be emphasized in this paper is the location-allocation problem among rectangular regions with bivariate uniform distributions.

This may be expressed as,

$$\text{(P')} \qquad \text{minimize} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{z_{ij} w_i}{A_i} \int_{a_i} \int_{b_i} |X_j - R_i|_{l_p} \, da_i db_i$$

$$\text{subject to: } \sum_{j=1}^{n} z_{ij} = 1 \qquad i = 1, \ldots, m$$

$$z_{ij} = 0, 1 \qquad \text{for all } i \text{ and } j$$

where: (a_i, b_i) = general coordinate location in region R_i
A_i = area of region R_i

and n, m, w_i, X_j, R_i and z_{ij} are as defined in (P).

Note that $\frac{1}{A_i}$ in (P') is just the bivariate uniform density function over R_i.

In Problems (P) and (P') , the decision variables are the z_{ij}'s-reflecting the allocation aspects of the problem and the X_j's-reflecting the location aspects of the problem.

The new facilities have an infinite capacity to serve the existing facilities. Thus, each existing facility will be allocated to and subsequently interact with only the closest new facility.

It is assumed throughout that the w_i's may represent either deterministic values or expected values of random variables. Also, the regions must be rectangular, but they may be overlapping.

3. RELATED WORK

There has been no previous work on the $L-A$ problem among regions. However, research on the deterministic version of the problem has revealed the complexities and computational burden involved in the solution of the $L-A$ problem.

In light of the difficulties associated with exact solution of the $L-A$ problem, heuristic algorithms are often employed. Cooper [5,6,7] developed various heuristic algorithms. Many of his initial algorithms used the assumption that all existing facilities were equally weighted; he used these results to develop heuristic for the case when the facilities are not equally weighted. Leamer [10] assumed customers were uniformly distributed over a plane and attempted to allocate them to the new facilities by dividing the plane into hexagonal areas.

Since the heuristics can not guarantee a specific proximity to optimality, exact algorithms have been developed with an attempt to alleviate the computational burden of the $L-A$ problem. Most algorithms have concentrated on the Euclidean metric. Bellman [4] was able to solve very small $L-A$ problems by transforming them into dynamic programming problems using quasilinearization as the transformation device. Kuenne and Soland [9] used a branch and bound algorithm to optimally solve the $L-A$ problem with Euclidean, great circle, and rectilinear distance metrics. Ostresh [12] worked on the Kuenne and Soland algorithm in an attempt to improve the bounding procedure. He did so for the case $n = 2$ using convexity results of Wendell and Hurter [15]. Love and Morris [11] considered the $L-A$ problem with rectilinear norm. Their exact algorithm features a reduction scheme where only possibly optimal sites for new facilities are considered. Recently, Sherali and Shetty [14] used a cutting plane algorithm to solve the $L-A$ problem with rectilinear norm.

Although these exact methods can guarantee optimality, there are limitations to the size of problem that can be solved in terms of computational time. Ostresh [12] reported solving problems of sizes $m = 23$, $n = 2$ and $m = 11$, $n = 4$ in respective CPU times of 23.26 sec and 10.28 sec on IBM 360/65. Kuenne and Soland's [9] largest reported problem was $m = 15$, $n = 4$ with CPU times for random weights and unit weights, respectively, of 82.7 sec and 54.2 sec on an IBM 360/91. Sherali and Shetty [14] solved a problem of size $m = 35$, $n = 2$ in 23.46 seconds on a CDC 6600. Finally, Love and Morris [11] reported solving a problem of size $m = 35$, $n = 2$ in one hour and 31 minutes of CPU time on a Univac 1110. Thus, computational burden seems to be a serious problem for exact solution methods.

4. A BRANCH AND BOUND APPROACH

The branch and bound approach developed by Kuenne and Soland [9] offers an optimal solution to the $L-A$ problem in reasonable computational time. Although Kuenne and Soland developed a solution for the deterministic problem, some of their results may be generalized and adapted to the form of the $L-A$ problem considered here. Some of the generalized results are discussed below.

The $L-A$ branch and bound algorithm is based on partitioning the set of all possible solutions to the location-allocation problem on the basis of the allocations of the existing facilities to the new facilities.

Any subset of solutions, denoted S, can be partitioned into at most n disjoint sets by considering the total number of ways a previously unallocated existing facility can enter the allocation scheme. Suppose that in S the allocated existing facilities have been assigned to k new facilities where $k \leqslant n$. An unallocated existing facility is chosen. If $k = n$, then S can be partitioned or separated into n subsets $S_1, S_2, \ldots, S_n$ where S_j is characterized by the assignment of the existing facility to new facility j. On the other hand, if $k < n$, then S may be partitioned into $K + 1$ subsets where S_j, $j = 1, 2, \ldots, k$ is as described above. The subset S_{k+1} is characterized by the assignment of the existing facility to a $(k + 1)$th new facility. This $(k + 1)$th new facility would have only one existing facility allocated to it.

After a node or subset A has been partitioned, a lower bound is computed for each partition or succeeding node j to help in fathoming the generated nodes. This bound is a lower bound on the objective function value that would be produced by any allocation scheme containing the allocations that have been made at this node j. The lower bound is a sum of two values. The first value is the cost of optimally locating the new facilities among the existing facilities that have been allocated; this is just a multifacility location problem. The second value is a lower bound on the cost of locating n new facilities among the unassigned existing facilities.

When the mth level is reached a complete allocation scheme has been developed, as each of the m existing facilities has been allocated to one of the n new facilities.

4.1 The Branching Rule

The branching rule is the criterion used to choose the unallocated existing facility at each level whose assignment will be considered as the basis for making the partition. Any rule may be used. For example, an unallocated existing facility could be chosen at random or the ith existing facility could be chosen as the branching facility at the ith level. However, an approach based on the properties of the problem may be more useful.

For this problem where the sum of weighted expected distances is to be minimized, the weighted expected distance from an existing facility to a new facility will be considered as a branching rule as a generalization of the results of Kuenne and Soland [9]. Considering only the minimum distance or maximum distance between an existing facility and all new facilities would disregard the size and variations of the expected distances between the existing facility and the new facilities.

The weighted expected distance between region i and new facility j is

$$\text{(1)} \qquad \frac{w_i}{A_i} \int_{b_{i_1}}^{b_{i_2}} \int_{a_{i_1}}^{a_{i_2}} (|x_j - a_i| + |y_j - b_i|)\, da_i\, db_i$$

where all parameters are defined as in (P) and (P').

This is equivalent to the following expression:

$$\text{(2)} \qquad \frac{w_i}{a_{i_2} - a_{i_1}} \int_{a_{i_1}}^{a_{i_2}} |x_j - a_i|\, da_i + \frac{w_i}{b_{i_2} - b_{i_1}} \int_{b_{i_1}}^{b_{i_2}} |y_j - b_i|\, db_i$$

Each expression in the sum may be computed independently. Hence, because of this separability there is an expected distance with respect to the x-coordinate and another with respect to the y-coordinate.

It may be shown that the expected distance from (x, y) to region i defined by $[a_{i_1}, a_{i_2}]\ x\ [b_{i_1}, b_{i_2}]$, where $x \notin (a_{i_1}, a_{i_2})$ and $y \notin (b_{i_1}, b_{i_2})$, is equivalent to the rectilinear distance from (x, y) to the midpoint of these intervals $\left(\frac{a_{i_1} + a_{i_2}}{2}, \frac{b_{i_1} + b_{i_2}}{2}\right)$. Another case is presented in Theorem 1.

These expected distances are used in both applying the branching rule and evaluating the objective function.

4.2 Upper and Lower Bounds

4.2.1 Bounds on the Objective Function

The objective function value associated with an arbitrary allocation scheme may serve as an upper bound. This upper bound may be improved by using a modification of Cooper's alternate location and allocation heuristic [6].

Consider the arbitrary allocation where existing facility i is allocated to new facility j where

$$j = \begin{cases} i \ (\mathrm{mod}\ n), & \text{if } j \text{ is not divisible by } n \\ n, & \text{otherwise.} \end{cases}$$

By this definition existing facility n would be allocated to new facility n, but existing facility $n + 1$ would be allocated to new facility 1.

The location problem for this allocation is solved and the objective function value computed. This is an upper bound on the optimal solution value. The upper bound is tested for improvement by reallocating each existing facility to the new facility whose weighted expected distance from the former facility is a minimum. After the reallocations are made, the location problems are again solved and a new objective function value computed. If the new objective function value is equal to the old objective function value, iterations cease. Otherwise, the reallocations start again. This heuristic may be iterated until no improvement is made or until a convergence criterion is met. The best objective function value from this heuristic becomes the upper bound on the optimal objective function value. The minimum expected cost of serving a region is established in the next theorem.

THEOREM 1: The minimum expected cost, T_i, of serving region i from a point within i, is $\frac{w_i}{4}\ (a_{i_2} - a_{i_1} + b_{i_2} - b_{i_1})$ where region i is defined as $[a_{i_1}, a_{i_2}]\ x\ [b_{i_1}, b_{i_2}]$.

PROOF: The expected cost of serving region i from (x, y) a point within i is

$$(3) \qquad f(x, y) = w_i \left[\frac{(a_{i_2} - x)^2 + (a_{i_1} - x)^2}{2(a_{i_2} - a_{i_1})} + \frac{(b_{i_2} - y)^2 + (b_{i_1} - y)^2}{2(b_{i_2} - b_{i_1})}\right].$$

When the partial derivatives of (3) are set to 0, the solution is

$$(x^*, y^*) = \left(\frac{a_{i_2} + a_{i_1}}{2}, \frac{b_{i_2} + b_{i_1}}{2}\right).$$

which yields a minimum. Thus,

$$f(x^*, y^*) = \frac{w_i}{4}\,(a_{i_2} - a_{i_1} + b_{i_2} - b_{i_1}) = T_i.$$

The lower bound to the objective function may then be found by:

$$\text{(4)} \qquad l.b. = \sum_{i=1}^{m} T_i.$$

4.2.2 A Lower Bound for Each Node

Computing a lower bound is a two part process. The first part is solving the location problem for the allocated existing facilities and computing the corresponding new facility. The second part involves underestimating the expected cost of locating the n new facilities among the unallocated existing facilities.

In order to develop the second expression, consider two unallocated regions R_1 and R_2. Suppose that both are to be served by the same new facility $X = (x, y)$. The expected cost of serving these two regions is:

$$\text{(5)} \qquad f(X) = w_1 E[|x - a_1| + |y - b_1|] + w_2 E[|x - a_2| + |y - b_2|]$$

where (a_i, b_i) are random variables representing the points located in region i. This expression can be considered the sum of the expected costs of serving the regions along the x-coordinate and the expected cost of serving the regions along the y-coordinate. These expressions are independent and each one-dimensional case may be considered separately.

Notice that when the x-coordinate is considered, then the expected cost is

$$\text{(6)} \qquad f(x) \geqslant \min\{w_1, w_2\}\,(E[|x - a_1|] + E[|x - a_2|]).$$

Let a_1 and a_2 assume any values where $a_1 < a_2$ and consider the relative position of x. By the triangle inequality,

$$\text{(7)} \qquad |x - a_1| + |x - a_2| \geqslant |a_1 - a_2|.$$

Since a_1 and a_2 are random variables, then

$$\text{(8)} \qquad E[|x - a_1|] + E[|x - a_2|] \geqslant E[|a_1 - a_2|].$$

Substituting (8) into (6), a lower bound is produced:

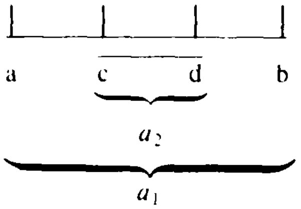
$$\text{(9)} \qquad f(x) \geqslant \min\{w_1, w_2\}\, E[|a_1 - a_2|].$$

Thus, (9) is an appropriate lower bound, where $E[|a_1 - a_2|]$ represents the expected distance between regions 1 and 2 along the x-coordinate.

$$\text{(10)} \qquad E[|a_1 - a_2|] = \int_{a_{1_1}}^{a_{1_2}} \int_{a_{2_1}}^{a_{2_2}} |u - v|\, du\,dv.$$

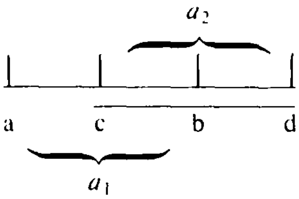
The integral in (10) may be evaluated for three cases. For ease in reading, let a represent a_{1_1}, b represent a_{1_2}, c represent a_{2_1}, and d represent a_{2_2} (the second interval a_2 is underlined).

CASE I. $a < c < d < b$



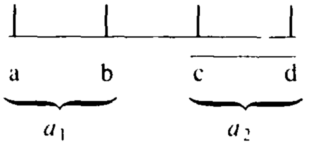
$$(11)\qquad E[|a_1 - a_2|] = \frac{(a^2 + b^2)(d - c) - (a + b)(d^2 - c^2) + \frac{2}{3}(d^3 - c^3)}{2(b - a)(d - c)}.$$

CASE II. $a < c < b < d$



$$(12)\quad E[|a_1 - a_2|] = \frac{(b - c)[(a^2 + c^2) - (a + b)(b + c)] + \frac{2}{3}(b^3 - c^3) + (d - b)(b - a)(d - c)}{2(b - a)(d - c)}.$$

CASE III. $a < b < c < d$



$$(13)\qquad E[|a_1 - a_2|] = \frac{(d^2 - c^2)(b - a) - (d - c)(b^2 - a^2)}{2(b - a)(d - c)}.$$

From (13) it can be shown that if the two regions R_1 and R_2 have nonoverlapping intervals, then the expected distance between the two is just $\left|\frac{d + c - b - a}{2}\right|$.

Thus, for any two rectangular regions R_i and R_j, the expression

$$(14)\qquad \min\{w_i, w_j\}(E[|a_i - a_j| + |b_i - b_j|])$$

can be computed as an underestimate of the expected cost of serving these two regions with the same new facility (see [6]).

Thus, Equation (14) is the building block for forming lower bounds. If there are p unallocated existing facilities, then there are $1/2p(p - 1)$ different realizations of Equation (14). Assume that all the expressions are placed in ascending order and let q_i be the ith term in this progression. Compute T_j for $j = 1, \ldots, P$, where T_j is defined as in (4), and arrange these expressions in ascending order. Let r_i be the ith term in this progression.

To underestimate the expected cost of allocating n new facilities among p existing facilities, the various combinations of allocations should be studied. For example, if $p \leqslant n$, then a new facility should be assigned to each of the p existing facilities. An underestimate of this cost would be the sum of all p of the r_i terms. This would follow since r_i represents a minimum expected cost for serving a region from a point in the region.

Another example is the case where $p = n + 4$. In this case, there are five possible combinations: four new facilities are allocated two existing facilities, all others are allocated one; one new facility is allocated three existing facilities, two are allocated two, and the others are allocated one; one new facility is allocated four existing facilities, one is allocated two, and all others are allocated one; two new facilities are allocated three existing facilities apiece, and all others are allocated one; and finally one new facility is allocated five existing facilities, and all others are allocated one.

Table 1 displays all lower bounds for these combinations for different values of $p - n$.

TABLE 1 — *Lower Bounds for Locating n New Facilities Among p Rectangular Regions*

Value of $(p - n)$	Lower Bound
0 or less	$\sum_{i=1}^{m} r_i$
1	$q_1 + \sum_{i=1}^{n-1} r_i$
2	$\min \left\{ q_1 + q_2 + \sum_{i=1}^{n-2} r_i,\ 1/2(q_1 + q_2 + q_3) + \sum_{i=1}^{n-1} r_i \right\}$
3	$\min \left\{ q_1 + q_2 + q_3 + \sum_{i=1}^{n-3} r_i,\ q_1 + 1/2(q_2 + q_3 + q_4) + \sum_{i=1}^{n-2} r_i, \right.$ $\left. \frac{1}{3}\ (q_1 + \ldots + q_6) + \sum_{i=1}^{n-1} r_i \right\}$
4	$\min \left\{ q_1 + \ldots q_4 + \sum_{i=1}^{n-4} r_i,\ q_1 + q_2 + 1/2(q_3 + q_4 + q_5) + \sum_{i=1}^{n-3} r_i, \right.$ $q_1 + \frac{1}{3}\ (q_2 + \ldots + q_7) + \sum_{i=1}^{n-2} r_i,\ 1/2(q_1 + \ldots + q_6) + \sum_{i=1}^{n-2} r_i,$ $\left. \frac{1}{4}\ (q_1 + \ldots + q_{10}) + \sum_{i=1}^{n-1} r_i \right\}$
5	$\min \left\{ q_1 + \ldots + q_5 + \sum_{i=1}^{n-5} r_i,\ q_1 + q_2 + q_3 + 1/2(q_4 + q_5 + q_6) + \sum_{i=1}^{n-4} r_i, \right.$ $q_1 + q_2 + \frac{1}{3}\ (q_3 + \ldots + q_8) + \sum_{i=1}^{n-3} r_i,\ q_1 + 1/2(q_2 + \ldots + q_7) + \sum_{i=1}^{n-3} r_i,$ $q_1 + 1/4(q_2 + \ldots + q_{11}) + \sum_{i=1}^{n-2} r_i,\ 1/2(q_1 + q_2 + q_3) + \frac{1}{3}\ (q_4 + \ldots + q_9)$ $\left. + \sum_{i=1}^{n-2} r_i,\ \frac{1}{5}\ (q_1 + \ldots + q_{15}) + \sum_{i=1}^{n-1} r_i \right\}$

It is obvious that as $p - n$ becomes larger than five, the number of combinations to be considered also becomes large. Thus, a general lower bound will be used for values of $p - n$ greater than five.

THEOREM 2: A general lower bound on locating n new facilities among p rectangular regions is $1/2 \sum_{i=1}^{p-n} q_i$ where q_i is as defined above. The proof follows closely that in [6,9].

This general lower bound is well-suited for the cases when $p - n$ is large. These cases will be levels 1, 2, ..., $m - 5$ of the tree. At these levels, the possibility of fathoming nodes is not as great as at the other levels. This is because only a few existing facilities have been allocated, and the partial objective function value used in computing the lower bound will be far from the optimum. A tight lower bound would then involve considering all possible combinations of the unallocated facilities. To hasten the tree search, the general lower bound is used to quickly compute the lower bound and move to the next level.

On the other hand, in the last $n + 5$ levels of the tree, enough facilities have been allocated to identify unprofitable allocation schemes. Here the tighter lower bounds given in Table 1 should be used to fathom as many nodes as possible.

5. THE LOCATION-ALLOCATION BRANCH AND BOUND ALGORITHM (LABB)

In this section the complete branch and bound algorithm for the location-allocation problem is given.

The input parameters are

N = number of new facilities
M = number of existing regions
$x1(I)$ and $x2(I)$ = left and right endpoints, respectively, of region $I(R(I))$ along x-axis
$y1(I)$ and $y2(I)$ = lower and upper endpoints, respectively, along y-axis.
$w(I)$ = interaction cost for region I.

The parameters for computing bounds on the optimum value of the objective function are:

$\bar{z}$ = upper bound on optimum
$\underline{z}$ = lower bound on optimum
FX = current least upper bound on optimum
F = objective function value to be compared with FX
ϵ = stopping criterion for alternate heuristic ($\epsilon > 0$).

The parameters for computing the branching facility are:

L = current level
J_l = index of branching facility chosen at level L
IJ_l = set of indices of unallocated facilities at level L
$AED(I)$ = vector of average expected distances from region I to all other regions.
$AX(I)$ = vector of average distance of region I to the new facilities that have been currently located.

The parameters for creating and fathoming new nodes are:

KL = number of new nodes to be created at current level
$NODE$ = counter for nodes created
ND = node number of the last node created at previous level
$IP(L)$ = the new facility the branching facility at level L was allocated to according to the node that was partitioned at level L.
NL = number of new facilities at previous level
$XX(J)$ = current location of new facility J
$XLB(I)$ = lower bound at node I
$Q(I)$ = the ith smallest value of $\min\{w(j),\ w(k)\}E[|R(j) - R(k)|]$ for all $j < k$
$R(I)$ = the ith smallest value of $.25w(j)[x2(j) - x1(j) + y2(j) - y1(j)]$.

STEP 0. Initialize the input parameters. (Compute upper and lower bounds on optimum.)

STEP 1. Let $FX = \infty$.

STEP 2. Arbitrarily allocate region I to new facility $I - N\left\lfloor \frac{I-1}{N} \right\rfloor$.

STEP 3. Solve the single facility location problem for all new facility $XX(j)$, $j = 1, \ldots, n$ among the regions allocated to new facility j.

STEP 4. Evaluate F, the objective function value of the $L-A$ problem, for the results of Step 3.

STEP 5. If $FX - F > \epsilon$, then replace FX with F. Otherwise, go to 7.

STEP 6. For $I = 1, \ldots, M$, compute $\min_j\{w(I) \cdot E[|XX(j) - R(I)|]\}$; let k be that facility with the minimum expected value. Reallocate region I to new facility k. Go to 3.

STEP 7. Let $\bar{z} = FX$.

STEP 8. Compute $\underline{z} = .25 \sum_{I=1}^{M} w(I) \cdot [x2(I) - x1(I) + y2(I) - y1(I)]$.

STEP 9. If $\bar{z} = \underline{z}$, stop. Go to 31.

(Initialize for level 1)

STEP 10. $L = 1$.

STEP 11. For $I = 1, \ldots, M$, compute $AED(I) = \sum_{k=1}^{m} E[|R(I) - R(k)|]$.

STEP 12. Let $i_1 = \max_I AED(I)$. $II_1 = \{1, 2, \ldots M\} - i_1$.

STEP 13. Let $NODE = 1$. Assign region i_1 to new facility 1. Solve the location problem for $XX(1)$. Let $NL = 1$. Let $IP(1) = 1$.

(Advance to next level)

STEP 14. Let $L = L + 1$.

(Compute Branching Facility)

STEP 15. Compute $AX(I) = \frac{1}{NL} \sum_{II=1}^{NL} E[|R(I) - XX(II)|]$ for $I \epsilon IJ_{l-1}$.

STEP 16. Let $j_l = \max_I AX(I)$ and let $IJ_l = IJ_{l-1} - j_l$.

(Create New Nodes)

STEP 17. Let $KL = \min(L, N)$. Let $ND = NODE$.

STEP 18. Create KL new nodes $ND + 1, \ldots, ND + KL$ by allocation region j_l to new facility $1, \ldots, KL$, respectively. Let $NODE = ND + KL$.

(Compute Lower Bounds on Nodes)

STEP 19. For node $I = ND + 1, \ldots, ND + KL$, solve the location problem for the partial allocation scheme: region j_l allocated to new facility $I - ND$; j_k allocated to $IP(k)$, $k = L - 1, L - 2, \ldots, 1$. Denote the objective function value $LB(I)$.

STEP 20. Compute the vectors $Q(I)$ and $R(I)$ using regions J, $J \epsilon IJ_l$.

STEP 21. If $M - L - N > 5$, let $Qx = 1/2 \sum_{I=1}^{M-L-N} Q(I)$. If $M - L - N \leqslant 5$, compute the lower bound for the value $M - L - N$ as given in Table 1. Denote this value Qx.

STEP 22. Let $LB(I) = LB(I) + Qx$, $I = ND + 1, \ldots, ND + k$. If $LB(I) \geqslant \Xi$, fathom node I.

STEP 23. Among the unfathomed nodes in 22, choose I^* as the value of I such that $LB(I^*) = \min_I LB(I)$. If all nodes are fathomed, go to 27.

STEP 24. Let $IP(L) = I^* - ND$.

STEP 25. If $L < M$, set NL and $XX(j)$ $j = 1, \ldots, NL$ equal to the values found for I^* in Step 19 and go to 14.

STEP 26. If $L = M$, compare $LB(I^*)$ to Ξ. If $LB(I^*) < \Xi$ then $\Xi = LB(I^*)$. Fathom the newly created nodes at level M.

(Backtracking Procedure)

STEP 27. Let $L = L - 1$. If $L = 1$, stop. Go to 31.

STEP 28. Consider all nodes I at level L that are unfathomed and have not been partitioned such that their allocation scheme includes j_{L-1} allocated to $IP(L - 1)$. If there is a node I such that $LB(I) < \Xi$, go to 29. Otherwise, go to 27.

STEP 29. Choose I^* such that $LB(I^*) = \min_I LB(I)$ where I are the active nodes identified in 27. Let LL denote the new facility j_l was allocated to at i^*. $IP(L) = LL$. Let NL and $XX(j)$ become the appropriate values found in Step 19 for I^*.

STEP 30. Go to Step 14

STEP 31. The optimal allocation scheme is the one associated with $\bar{z}$, the optimal objective function value.

6. VERIFICATION OF THE ALGORITHM

LABB, was coded in Fortran IV. The code was verified using an example problem presented in Figure 1 where the w_i's are 2,1,2,2,1, respectively, for the five regions.

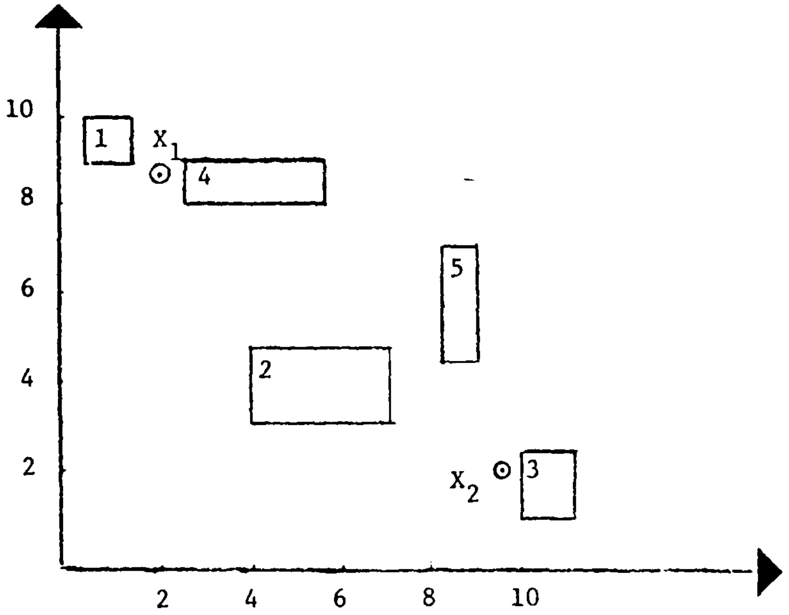


FIGURE 1. A graph of an example problem

Both manual computation and the code produced the optimal allocation scheme to be: X_1 serves regions 1 and 4 and X_2 serves regions 2, 3 and 5. The two new facilities X_1 and X_2 were located at (2.5,9) and (9.5,1.5), respectively. The optimal objective function value was 18.5.

The same problem with a centroid approximation produced a different allocation scheme: X_1 serves regions 3 and 5 and X_2 serves regions 1, 2 and 4. The new facilities were located at the points (4,8.5) and (9.5,1.5), the centroids of regions 4 and 3, respectively. These locations used in the objective function involving the rectangular regions produced a value of 37, a 100 percent increase over the value for the optimal locations.

The impact of the sensitivity of the rectilinear distance metric to the centroid approach on the location-allocation problem is serious; it has produced a nonoptimal allocation scheme and inferior locations for the new facilities. As in the multifacility location problem, the centroid approach does not even offer a good approximation to the solution of the location-allocation model.

7. COMPUTATIONAL RESULTS

The computational results given in this section represent experience with the branch and bound algorithm (LABB) for rectangular regions using a rectilinear distance metric. The problems were randomly generated from uniform distributions. All $w(I)$'s were generated from a uniform [0,10] distribution. The $x1(I)$'s, $x2(I)$'s, $y1(I)$'s and $y2(I)$'s were each generated from a uniform [0,100] distribution. All problems were run on an IBM 370/158J computer. The results are summarized in Table 2.

TABLE 2 — *Computational Results for Location-Allocation Problems where n = 2, 3, and 4*

m	No. of Problems	Average CPU Time (seconds)	Average No. of Nodes Created	Average Maximum No. of Active Nodes	Average Optimal Node
n = 2					
5	2	.585	11	1	10
6	3	1.01	27.3	2.7	19.7
7	4	1.00	25.75	3.5	20
9	3	1.99	94.33	4.67	79
11	3	5.37	240.3	8	119
15	2	8.55	330	11	219
20	2	11.33	332	18	39
25	2	19.08	349	23	69
30	1	33.02	517	28	59
35	1	51.15	541	33	69
n = 3					
6	3	1.2	50	3.3	32
7	3	1.54	86	5.33	68
9	3	3.76	232	10.33	59
11	2	4.51	272	14	211.5
15	2	7.28	412.5	23	188
20	3	11.2	411	35.7	62.3
25	2	15.44	—	45	72
30	2	26.37	564	55	87
35	1	37.23	543	65	351
n = 4					
7	3	2.73	129.33	7.3	87.33
9	3	3.01	223	11.3	118
11	2	4.8	316	23	38
15	2	6.73	416	36	54
20	4	11.77	586	48	74
25	1	13.26	[illegible]	66	94

Not surprisingly, the required computational time reflects the average number of nodes created which, in turn, is a function of the size of the problem and the number of active nodes. For the problems worked, no computational time was over one minute.

In each of the cases of m for $n = 2$, the optimal allocation was examined to determine what percentage of optimum was achieved by the lower bound at each level. In cases where m was large, the general lower bound was used at the first $m - 6$ levels. The lower bound improved rapidly from level to level; a typical improvement was ten percent of optimum. Usually, at the $m - 6$th level, the lower bound was within 85-90 percent of optimum. Thus, the switch to the combinatorial lower bounds for the last five levels represented less improvement from level to level, but convergence occurred rapidly.

The computational results in Table 2 indicate that the LABB algorithm obtains an optimal solution for the $L-A$ problem with rectangular regions in very reasonable time.

8. SUMMARY

In this paper the location-allocation problem for existing facilities uniformly distributed over rectangular regions was considered. Previous works dealing with $L-A$ systems were discussed, and the properties of the problem were developed. These properties indicated that developing the optimal allocation scheme was the most important step in optimally solving the $L-A$ problem.

Computational results indicated that the exact algorithm (LABB) could obtain the optimal solution for large problems in a reasonable time.

The branch and bound method (LABB) may be applied to location-allocation problems with probability distributions on existing facilities other than uniform. Since the branch and bound methods generate optimal allocation schemes no matter what type of objective function is used, the only difference would be the way the location problems are solved at each node. Solution techniques using other probability distributions are developed by Aly [1], Katz and Cooper [8] and Wesolowsky [17]. It would be expected that the computational times to solve these related problems would be similar to the times for the uniform distribution with adjustments made on the basis of the speed of the individual solution technique.

A $L-A$ problem may have constraints on the allocation scheme, on the locations of the new facilities, or on both. In these cases, the constraints can be used as an additional test at each node as a basis for fathoming the node.

REFERENCES

[1] Aly, A.A., "Probabilistic Formulation of Some Facility Location Problems," unpublished Ph.D. dissertation, Virginia Polytechnic Institute, Blacksburg, Va. (1975).

[2] Aly, A.A. and White, J.A., "Probabilistic Formulation of the Multifacility Weber Problem," Naval Research Logistics Quarterly, 25, 531-547 (1978).

[3] Aly, A.A. and Steffen, A.E., "Multifacility Location Problem Among Rectangular Regions," Working paper, School of Industrial Engineering, University of Oklahoma, Norman, Okla. (1978).

[4] Bellman, R.E., "An Application of Dynamic Programming to Location-Allocation Problems," SIAM Review, 7, 126-128 (1965).

[5] Cooper, L., "Location-Allocation Problems," Operations Research, *11*, 331-334 (1963)

[6] Cooper, L., "Heuristic Methods for Location-Allocation Problems," SIAM Review, *6*, 37-53 (1964).

[7] Cooper, L., "Generalized Locational Equilibrium Models," Journal of Regional Science, 7, 1-17 (1967).

[8] Katz, I.N. and Cooper, L., "An Always Convergent Numerical Scheme for a Random Locational Equilibrium Problem," SIAM Journal of Numerical Analysis, *17*, 683-692 (1974).

[9] Kuenne, R.E. and Soland, R.M., "Exact and Approximate Solutions to the Multisource Weber Problem," Mathematical Programming, *3*, 193-209 (1972).

[10] Leamer, E.E., "Locational Equilibrium," Journal of Regional Science, *8*, 229-242 (1968).

[11] Love, R.F. and Morris, J.G., "A Computation Procedure for the Exact Solution of Location-Allocation Problems with Rectangular Distances," Naval Research Logistics Quarterly, *22*, 441-453 (1975).

[12] Ostresh, L.M., "An Efficient Algorithm for Solving the Two Center Location-Allocation Problem," Journal of Regional Science *15*, 209-216 (1975).

[13] Rushton, G., Goodchild, M.F. and Ostresh, L.M., eds., *Computer Programs for Location-Allocation Problems*, Monograph No. 6, Department of Geography, University of Iowa, Iowa City, Ia. (1973).

[14] Sherali, A.D. and Shetty, C.M., "The Rectilinear Distance Location-Allocation Problem," AIIE Transactions, *9*, 136-143 (1977).

[15] Wendell, R.E. and Hurter, A.P., "Location Theory, Dominance and Convexity," Operations Research, *21*, 314-320 (1973).

[16] Wesolowsky, G.O. and Love, R.F., "Location of Facilities with Rectangular Distances Among Point and Area Destination," Naval Research Logistics Quarterly, *18*, 83-90 (1971).

[17] Wesolowsky, G.O., "The Weber Problem with Rectangular Distances and Randomly Distributed Destinations," Journal of Regional Science, *17*, 53-59 (1977).

AN ITERATIVE ALGORITHM FOR THE MULTIFACILITY MINIMAX LOCATION PROBLEM WITH EUCLIDEAN DISTANCES

Christakis Charalambous

Department of Electrical Engineering
Concordia University
Montreal, Quebec, Canada

ABSTRACT

An iterative solution method is presented for solving the multifacility location problem with Euclidean distances under the minimax criterion. The iterative procedure is based on the transformation of the multifacility minimax problem into a sequence of squared Euclidean minisum problems which have analytical solutions. Computational experience with the new method is also presented.

1. PROBLEM FORMULATION

To formulate the problem, let us suppose that m existing facilities are located at known points $(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)$ and n new facilities are to be located at points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The cost

$$\text{(1a)} \qquad f_{ij}(x_i, y_i) = w_{ij}[(x_i - a_j)^2 + (y_i - b_j)^2]^{1/2}, \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{matrix}$$

is incurred due to travel between new facility i and existing facility j for all i and j (w_{ij} is a nonnegative weight) and the cost

$$\text{(1b)} \qquad g_{lk}(x_l, y_l, x_k, y_k) = v_{lk}[(x_l - x_k)^2 + (y_l - y_k)^2]^{1/2}, \qquad \begin{matrix} l = 1, 2, \ldots, n-1 \\ k = l+1, \ldots, n \end{matrix}$$

is incurred due to travel between new facilities l and k for all $l < k$ (v_{lk} is a nonnegative weight).

From (1a) and (1b) we can see that the maximum cost incurred due to movement between facilities is:

$$\text{(2)} \qquad F(x, y) = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m \\ 1 \leq l < k \leq n}} \{f_{ij}(x_i, y_i), g_{lk}(x_l, y_l, x_k, y_k)\}$$

where

$$x = [x_1, x_2, \ldots, x_n]^T, \quad y = [y_1, y_2, \ldots, y_n]^T.$$

The multifacility Euclidean minimax facility location problem is to find (x, y) which minimizes $F(x, y)$. The new facilities might be helicopter bases, transmitting stations where it is desired to minimize the necessary signal, detection stations or civil defense sirens. An interesting book in this area is that given in reference [8].

One main characteristic of the objective function $F(x, y)$ is that it has discontinuous partial derivatives at points where two or more of the functions $f_{ii}(x_i, y_i)$, $g_{lk}(x_l, y_l, x_k, y_k)$ are equal to $F(x, y)$. Various algorithms have been proposed for solving the general minimax problem, some of the most relevant of which are due to Charalambous and Conn [2], Charalambous [1], Dem'yanov and Malozemov [4], Madsen [11], and Dutta and Vidyasagar [6]. More specialized algorithms for the minimax location problems were published by Chatelon, Hearn and Lowe [3], Drezner and Wesolowsky [5], Elzinga, Hearn and Randolph [7], and Love et al. [10].

In this paper we present a simple algorithm to minimize $F(x, y)$. The original problem is transformed into a sequence of unconstrained squared Euclidean minisum problems which have analytical solutions. The resulting method is efficient and easy to implement on a computer. Numerical results are presented which illustrate the usefulness of the new method to the multifacility location problem.

2. THEORETICAL RESULTS

LEMMA 1: The functions $f_{ii}(x_i, y_i)$ and $g_{lk}(x_l, y_l, x_k, y_k)$ as defined in (1a) and (1b) respectively, are convex functions.

PROOF: See [9].

LEMMA 2: The function $F(x, y)$ is continuous and convex.

PROOF: This follows from the fact that each $f_{ii}(x_i, y_i)$ and $g_{lk}(x_l, y_l, x_k, y_k)$ are continuous and convex functions (see for example [4]).

Let $p_{ii}(x_i, y_i)$ and $q_{lk}(x_l, y_l, x_k, y_k)$ be the following $2n$-dimensional column vectors:

$$\text{(3a)}\quad p_{ii}(x_i, y_i) = \begin{array}{r|c|l} & 0 & \\ & \vdots & \\ & 0 & \\ (n) & x_i - a_i & \leftarrow (i) \\ & 0 & \\ & \vdots & \\ & 0 & \\ \hline & 0 & \\ & \vdots & \\ & 0 & \\ (n) & y_i - b_i & \leftarrow (n+i) \\ & 0 & \\ & \vdots & \\ & 0 & \end{array} ,\qquad \text{(3b)}\quad q_{lk}(x_l, y_l, x_k, y_k) = \begin{array}{|c|l} 0 & \\ \vdots & \\ (x_l - x_k) & \leftarrow (l) \\ \vdots & \\ (x_k - x_l) & \leftarrow (k) \\ 0 & \\ \hline 0 & \\ \vdots & \\ (y_l - y_k) & \leftarrow (n+l) \\ \vdots & \\ (y_k - y_l) & \leftarrow (n+k) \\ 0 & \end{array}$$

All the elements of p_{ij} are equal to zero except the ith and the $(n + i)$th, and all the elements of q_{lk} are equal to zero except the lth, kth, $(n + l)$th and the $(n + k)$th. Also note the $p_{ij}(x_i, y_i)$ and $q_{lk}(x_l, y_l, x_k, y_k)$ are the gradient vectors of the following functions

$$\frac{1}{2}\,[(x_i - a_j)^2 + (y_i - b_j)^2]$$

and

$$\frac{1}{2}\,[(x_l - x_k)^2 + (y_l - y_k)^2]$$

with respect to (x, y) respectively.

THEOREM 1 (Necessary and sufficient conditions for optimality): The necessary and sufficient conditions for the point (x^*, y^*) to be a minimum point for the function $F(x, y)$ are that there exist *nonnegative* multipliers $\lambda^*_{ij}(1 = 1, 2, \ldots, n,\ \ j = 1, 2, \ldots, m)$, $\mu^*_{lk}(l = 1, 2, \ldots, n - 1, k = l + 1, \ldots, n)$ such that

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda^*_{ij} \frac{w_{ij}^2}{f_{ij}(x_i^*, y_i^*)}\, p_{ij}(x_i^*, y_i^*)$$

$$+ \sum_{l=1}^{n-1}\sum_{k=l+1}^{n} \mu^*_{lk} \frac{v_{lk}^2}{g_{lk}(x_l^*, y_l^*, x_k^*, y_k^*)}\, q_{lk}(x_l^*, y_l^*, x_k^*, y_k^*) = 0 \tag{4a}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda^*_{ij} + \sum_{l=1}^{n-1}\sum_{k=l+1}^{n} \mu^*_{lk} = 1 \tag{4b}$$

$$\lambda^*_{ij}(F(x^*, y^*) - f_{ij}x_i^*, y_j^*)) = 0, \quad \begin{array}{l} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{array} \tag{4c}$$

$$\mu^*_{lk}(F(x^*, y^*) - g_{lk}(x_l^*, y_l^*, x_k^*, y_k^*)) = 0 \quad \begin{array}{l} l = 1, 2, \ldots, n - 1 \\ k = l + 1, \ldots, n \end{array} \tag{4d}$$

PROOF: The proof follows directly from the Kuhn-Tucker conditions for optimality for this problem or from the Corollary of Theorem 3.2 of Dem'yanov and Malozemov [4]. Note that $\lambda^*_{ij} = \mu^*_{lk} = 0$ for the functions $f_{ij}(x_i, y_i)$ and $g_{lk}(x_l, y_l, x_k, y_k)$ which are less than $F(x^*, y^*)$ at (x^*, y^*), i.e., for those functions which are *not active* at the solution (x^*, y^*). The λ^*_{ij} and μ^*_{lk} are called minimax multipliers. Also not that since $f_{ij}(x_i^*, y_i^*) = g_{lk}(x_l^*, y_l^*, x_k^*, y_k^*) = F(x^*, y^*)$ from (4c) and (4d)) for corresponding $\lambda^*_{ij} \neq 0$ and $\mu^*_{lk} \neq 0$, the denominators for all terms in the summations in (4a) can be replaced by 1.

The possibility that some $f_{ij}(x_i^*, y_i^*)$ or $g_{lk}(x_l^*, y_l^*, x_k^*, y_k^*) = 0$ can occur. In this case replacing the offending term by $\delta > 0$ will not change the optimality conditions since the associated Lagrange multiplier will be zero for nontrivial problems.

Consider now the following problem (Euclidean-distance minisum location problem).

For *given* nonnegative values of $\lambda_{ij} = \bar{\lambda}_{ij}$ and $\mu_{lk} = \bar{\mu}_{lk}$

$$\underset{(x,y)}{\text{minimize}}\ \Phi(x, y, \bar{\lambda}, \bar{\mu})$$

where

$$\Phi(x, y\ \bar{\lambda}, \bar{\mu}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \bar{\lambda}_{ij} w_{ij}^2[(x_i - a_j)^2 + (y_i - b_j)^2]$$
$$+ \frac{1}{2}\sum_{l=1}^{n-1}\sum_{k=l+1}^{n} \bar{\mu}_{lk} v_{lk}^2[(x_l - x_k)^2 + (y_l - y_k)^2] \tag{5}$$

λ_{ij} and μ_{lk} are going to be called estimates for the minimax multipliers.

Let

(6a) $$\bar{w}_{ij} = \bar{\lambda}_{ij} w_{ij}^2 (\geqslant 0) \quad \begin{array}{l} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{array}$$

(6b) $$\bar{v}_{lk} = \bar{\mu}_{lk} v_{lk}^2 (\geqslant 0) \quad \begin{array}{l} l = 1, 2, \ldots, n-1 \\ k = l+1, \ldots, n. \end{array}$$

THEOREM 2: For given nonnegative values of $\lambda_{ij} = \bar{\lambda}_{ij}$ and $\mu_{lk} = \bar{\mu}_{lk}$ the function $\Phi(x, y, \bar{\lambda}, \bar{\mu})$ is convex.

PROOF: See [12].

THEOREM 3: If $\lambda_{ij} = \lambda^*_{ij}$ $(i = 1, 2, \ldots, n, j = 1, 2, \ldots, m)$, $\mu_{lk} = \mu^*_{lk}$ $(l = 1, 2, \ldots, n-1, k = l+1, \ldots, n)$, the minimax multipliers corresponding to a minimum point (x^*, y^*), then (x^*, y^*) is a global minimum point of $\Phi(x, y, \lambda^*, \mu^*)$.

PROOF: The gradient vector of $\Phi(x, y, \lambda^*, \mu^*)$ at the point (x^*, y^*) is:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda^*_{ij} w_{ij}^2 p_{ij}(x_i^*, y_i^*) + \sum_{l=1}^{n-1} \sum_{k=1+1}^{n} \mu^*_{lk} v_{lk}^2 q_{lk}(x_l^*, y_l^*, x_k^*, y_k^*) = 0$$

from Theorem 1.

Since $\phi(x, y, \lambda^*, \mu^*)$ is a convex function the results follows.

Therefore, if we knew λ^* and μ^*, we could obtain (x^*, y^*) in one step by minimizing $\Phi(x, y, \lambda^*, \mu^*)$. Since we do not know these optimum multipliers in advance we need to estimate them. Let $(\bar{x}, \bar{y})$ be a minimum point of $\Phi(x, y, \bar{\lambda}, \bar{\mu})$ for given values of $\bar{\lambda}$ and $\bar{\mu}$. Define

(7a) $$\bar{\lambda}^*_{ij} = \bar{\lambda}_{ij} f_{ij}(\bar{x}_i, \bar{y}_i)/s \quad \begin{array}{l} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{array}$$

(7b) $$\bar{\mu}^*_{lk} = \bar{\mu}_{lk} g_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k)/s \quad \begin{array}{l} l = 1, 2, \ldots, n-1 \\ k = l+1, \ldots, n \end{array}$$

where

(8) $$s = \sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}_{ij} f_{ij}(\bar{x}_i, \bar{y}_i) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}_{lk} g_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k).$$

Note that

(9a) $$\bar{\lambda}^*_{ij} \geqslant 0, \bar{\mu}^*_{lk} \geqslant 0$$

and

(9b) $$\sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk} = 1.$$

Also at the point $(\bar{x}, \bar{y})$ the gradient vector of $\Phi(x, y, \bar{\lambda}, \bar{\mu})$ must be zero. This gives us

(9c) $$\sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} \frac{w_{ij}^2}{f_{ij}(\bar{x}_i, \bar{y}_i)} p_{ij}(\bar{x}_i, \bar{y}_i) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk} \frac{v_{lk}^2}{g_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k)} q_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k) = 0$$

which when compared with (4a, 4b) of Theorem 1 suggests $\bar{\lambda}^*_{ij}$, $\bar{\mu}^*_{lk}$ are approximations to $\bar{\lambda}^*_{ij}$, $\bar{\mu}^*_{lk}$.

THEOREM 4: At a minimum point $(\bar{x}, \bar{y})$ of $\Phi(x, y, \bar{\lambda}, \bar{\mu})$ the following inequality holds:

$$F_l(\bar{x}, \bar{y}) \leqslant F(x^*, y^*) \leqslant F(\bar{x}, \bar{y})$$

where

$$F_l(\bar{x}, \bar{y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} f_{ij}(\bar{x}_i, \bar{y}_i) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk}\, g_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k)$$

and $\bar{\lambda}^*$ and $\bar{\mu}^*$ are as they were defined in (7a) and (7b) respectively.

PROOF: The right hand side inequality is obvious. Also

$$F(x, y) \geqslant F(x^*, y^*) = \min_{(x,y)} F(x, y)$$

$$= \min_{(x,y)} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} F(x, y) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk} F(x, y) \right]$$

(since $\bar{\lambda}^*$ and $\bar{\mu}^*$ satisfy (9b))

$$\geqslant \min_{(x,y)} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} f_{ij}(x_i, y_i) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk} g_{lk}(x_l, y_l, x_k, y_k) \right]$$

(from the definition of $F(x, y)$)

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\lambda}^*_{ij} f_{ij}(\bar{x}, \bar{y}) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \bar{\mu}^*_{lk} g_{lk}(\bar{x}_l, \bar{y}_l, \bar{x}_k, \bar{y}_k), \qquad \text{(from (9c))}$$

$$= F_l(\bar{x}, \bar{y}).$$

3. THE ALGORITHM

The above theoretical results suggest the following algorithm:

STEP 1: Set $r = 1$

$$\lambda_{ij}^{(r)} = 1, \quad i = 1, 2, \ldots, n \qquad j = 1, 2, \ldots, m.$$
$$\mu_{lk}^{(r)} = 1, \quad l = 1, 2, \ldots, n-1 \qquad k = l + l, \ldots, n.$$

STEP 2: Find the minimum point $(x^{(r)}, y^{(r)})$ of $\Phi(x, y, \lambda^{(r)}, \mu^{(r)})$. (See later for details).

STEP 3: At the point $(x^{(r)}, y^{(r)})$ calculate f_{ij} and g_{lk} and update λ_{ij} and μ_{lk} as follows: Set

$$s^{(r)} = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij}^{(r)} f_{ij}(x_i^{(r)}, y_i^{(r)}) + \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \mu_{lk}^{(r)} g_{lk}(x_l^{(r)}, y_l^{(r)}, x_k^{(r)}, y_k^{(r)})$$

$$\lambda_{ij}^{(r+1)} \leftarrow \frac{\lambda_{ij}^{(r)} f_{ij}(x_i^{(r)}, y_i^{(r)})}{s^{(r)}} \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m$$

$$\mu_{lk}^{(r+1)} \leftarrow \frac{\mu_{lk}^{(r)} g_{lk}(x_l^{(r)}, y_l^{(r)}, x_k^{(r)}, y_k^{(r)})}{s^{(r)}} \quad l = 1, 2, \ldots, (n-1), \; k = l+1, \ldots, n.$$

STEP 4: Calculate

$$F_l(x^{(r)}, y^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij}^{(r+1)} f_{ij}(x_i^{(r)}, y_i^{(r)})$$

$$+ \sum_{l=1}^{n-1} \sum_{k=l+1}^{n} \mu_{lk}^{(r+1)} g_{lk}(x_l^{(r)}, y_l^{(r)}, x_k^{(r)}, y_k^{(r)}).$$

STEP 5: Stopping criterion: If $(F(x^{(r)}, y^{(r)}) - F_l(x^{(r)}, y^{(r)}))/F(x^{(r)}, y^{(r)}) < \epsilon$ stop; Otherwise set $r \leftarrow r + 1$ and go back to Step 2. (ϵ is a prescribed tolerance).

3.1 Finding the Optimum Solution of the Quadratic Function

For given nonnegative values of $\bar{\lambda}_{ij}$ and $\bar{\mu}_{lk}$ we want to find the minimizing point $(\bar{x}, \bar{y})$ for $\Phi(x, y, \bar{\lambda}, \bar{\mu})$. Let

(10a,b)

$$\bar{a} = \begin{bmatrix} \sum_{j=1}^{m} \bar{w}_{1j} a_j \\ \sum_{j=1}^{m} \bar{w}_{2j} a_j \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{m} \bar{w}_{nj} a_j \end{bmatrix} \qquad \bar{b} = \begin{bmatrix} \sum_{j=1}^{m} \bar{w}_{1j} b_j \\ \sum_{j=1}^{m} \bar{w}_{2j} b_j \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{m} \bar{w}_{nj} b_j \end{bmatrix}$$

(11)

$$A = \begin{bmatrix} \beta_1 & -\bar{v}_{12} & \cdot\ \cdot\ \cdot & -\bar{v}_{1n} \\ -\bar{v}_{12} & \beta_2 & \cdot\ \cdot\ \cdot & -\bar{v}_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \\ & & \cdot & \\ \cdot & \cdot & & \cdot \\ -\bar{v}_{1n} & -\bar{v}_{2n} & \cdot\ \cdot\ \cdot & \beta_n \end{bmatrix}$$

(12) $$\beta_i = \sum_{j=1}^{m} \bar{w}_{ij} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \bar{v}_{ij}, \quad i = 1, 2, \ldots, n$$

Then the optimum solution can be obtained by solving the two systems of equations (see, for example, [8]).

(13a,b) $$A\bar{x} = \bar{a} \quad \text{and} \quad A\bar{y} = \bar{b}.$$

Since for given nonnegative values of λ_{ij} and μ_{lk} the function $\Phi(x, y, \lambda, \mu)$ is convex, it follows that its Hessian matrix A is positive semi-definite. Also using the fact that A is symmetric we can write

$$A = LL^t \tag{14}$$

where L is an $n \times n$ lower triangular matrix.

This is called the Cholesky decomposition of A and requires about $n^3/6$ multiplications. By using (14) for each right hand side of (13), solve the following system to obtain $\bar{x}$ and $\bar{y}$.

$$Lp = \bar{a} \qquad Lq = \bar{b}$$
$$L^t\bar{x} = p \qquad L^t\bar{y} = q.$$

This requires about $2(n^2 + n)$ multiplications.

Note that if $\beta_i = 0$, then the ith row of A, the ith column of A, $\bar{a}_i$ and $\bar{b}_i$ are all equal to zero, and can be removed in solving for $\bar{x}$ and $\bar{y}$. In this case we have infinite solutions for the location (x_i, y_i) of the ith facility.

4. NUMERICAL EXAMPLES

We give a number of numerical examples to illustrate the usefulness of this approach to solving multifacility location minimax problems. For all the examples considered $\epsilon = 10^{-4}$. Computations were carried out at C 'ncordia University on CDC 64000 computer using single precision arithmetic. A user-oriented computer program written in Fortran IV implementing the above algorithm is available from ihe author.

EXAMPLE 1: Love, Wesolowsky and Kraemer [10], considered the problem where $n = 2$, $m = 5$ and $(a_1, b_1) = (39.12, 28.11)$, $(a_2, b_2) = (39.50, 28.28)$, $(a_3, b_3) = (37.88, 29.87)$, $(a_4, b_4) = (38.59, 27.03)$, $(a_5, b_5) = (38.38, 30.28)$, $v_{12} = 1$, and

$$W = (w_{ij}) = \begin{bmatrix} 1 & 4 & 4 & 4 & 1 \\ 4 & 1 & 1 & 1 & 4 \end{bmatrix}.$$

The results obtained by using the present approach are summarized below:

Results for Example 1

	Number of Iterations					
	1	2	6	10	30	40
$F(\bar{x}, \bar{y})$	6.5237	5.9768	5.9138	5.8738	5.8554	5.85496
$F_l(\bar{x}, \bar{y})$	4.3311	5.0112	5.5857	5.7367	5.8526	5.85439

Values of (λ_{ij}), (μ_{lk}), (f_{ij}), (g_{lk}) and $(\bar{x}, \bar{y})$ after 40 iterations:

$$\Lambda = (\lambda_{ij}) = \begin{bmatrix} 0. & 0.00047 & 0.49982 & 0.49965 & 0. \\ 0.00003 & 0. & 0. & 0. & 0.0003 \end{bmatrix}, \quad \mu_{12} = 0.$$

$$(f_{ij}(\bar{x}, \bar{y})) = \begin{bmatrix} 0.9476 & 5.1031 & 5.85466 & 5.85496 & 1.8355 \\ 4.58541 & 1.1831 & 1.1011 & 2.1709 & 4.58541 \end{bmatrix}, \quad g_{12} = 0.9052$$

$$\bar{x}_1 = 38.2356, \quad \bar{x}_2 = 38.7500 \qquad \bar{y}_1 = 28.4502, \quad \bar{y}_2 = 29.1950.$$

It can be seen that only functions f_{13} and f_{14} define the minimax function at the solution point and $\lambda_{ij} \to 0$, for all (i, j) except λ_{13} and λ_{14}. Also $\mu_{12} = 0$. In other words the λ_{ij} and

μ_{lk} corresponding to functions f_{ij} and g_{lk} that are not active at the minimax solution tend to zero, as it should be expected. Let

$$I_\lambda = \{(i, j) | f_{ij}(\bar{x}, \bar{y}) < 0.99 \quad F_l(\bar{x}, \bar{y}), \lambda_{ij} < 10^{-2}\}$$

$$I_\mu = \{(l, k) | g_{lk}(\bar{x}, \bar{y}) < 0.99 \quad F_l(\bar{x}, \bar{y}), \mu_{lk} < 10^{-2}\}$$

where $(\bar{x}, \bar{y})$ is the minimum point obtained at the end of the 40th iteration. If $(\bar{x}, \bar{y}) = (x^*, y^*)$, then the elements of I_λ and I_μ will correspond to functions which are not active at (x^*, y^*) and $\lambda^*_{ij} = 0$, $(i, j) \in I_\lambda$ $\mu^*_{lk} = 0$, $(l, k) \in I_\mu$. Also, if $(\bar{x}, \bar{y})$ is in the neighborhood of (x^*, y^*), then most likely the elements of I_λ and I_μ will correspond to functions which are not active at the solution.

By excluding from the problem the functions corresponding to the elements belonging in I_λ and I_μ, using the values of λ_{ij} and μ_{lk} obtained at the end of the 40th iteration for the remaining functions, the present algorithm reached the *exact* solution to the problem in *two* additional iterations. The final results obtained are summarized below. The method required 0.68 sec CPU time to reach the final results shown. From now on this additional part of the algorithm will be called *Phase 2*, and the original part of the algorithm where all functions are considered (algorithm in Section 3) will be called *Phase 1*:

Final Results for Example 1: $F(x^*, y^*) = F_l(x^*, y^*) = 5.85481$, $\lambda^*_{ij} = 0$. except $\lambda^*_{13} = \lambda^*_{14} = 0.5$, $\mu^*_{12} = 0$.

$$(f_{ij}(x^*, y^*)) = \begin{bmatrix} 0.9481 & 5.1055 & 5.85481 & 5.85481 & 1.8357 \\ 4.58541 & 1.1831 & 1.1011 & 2.1709 & 4.58541 \end{bmatrix}$$

$$g_{12}(x^*, y^*) = 0.9057$$

$$\left.\begin{array}{ll} x_1^* = 38.235 & y_1^* = 28.45 \\ x_2^* = 38.75 & y_2^* = 29.195 \end{array}\right\} \text{Exact solution.}$$

Since f_{13} and f_{14} are the *only* functions defining the minimax solution and both of them depend *only* on (x_1, y_1) (i.e., they are independent of the value of (x_2, y_2)). The value of (x_2^*, y_2^*) is *not unique*, but the value of (x_1^*, y_1^*) is unique. In fact, (x_2^*, y_2^*) is any point in the set

$$S^{(2)} = \bigcap_{j=1}^{6} S_j^{(2)}$$

where

$$S_j^{(2)} = \{(x, y) | w_{2j}[(x - a_j)^2 + (y - b_j)^2]^{1/2} \leqslant F(x^*, y^*)\}, \ j = 1, 2, \ldots, 5$$

$$S_6^{(2)} = \{(x, y) | v_{12}[(x - x_1^*)^2 + (y - y_1^*)^2]^{1/2} \leqslant F(x^*, y^*)\}.$$

The boundary of the set $S_j^{(2)}$ is a circle with center (a_j, b_j) and radius $F(x^*, y^*)/w_{2j}$, for $j = 1, 2, \ldots, 5$ and center (x_1^*, y_1^*) and radius $F(x^*, y^*)/v_{12}$ for $j = 6$. For this particular example the solution set $S^{(2)}$ for (x_2^*, y_2^*) is given by the intersection of the sets $S_1^{(2)}$ and $S_5^{(2)}$. This is illustrated in Figure 1. Since the value of (x_2^*, y_2^*) is not unique and our interest is on the minimax facility location problem it would be appropriate to choose the position of the second new facility such that the function

$$F_2(x_2, y_2) = \max_{1 \leqslant i \leqslant 5} \{f_{2i}(x_2, y_2), \quad g_{12}(x_2, y_2, x_1^*, y_1^*)\}$$

is minimized in the set $S^{(2)}$. The optimum solution to this problem occurs at the point C_2, and coincides with the minimum point obtained by using the present algorithm. In this case f_{21} and f_{25} define the minimax solution.

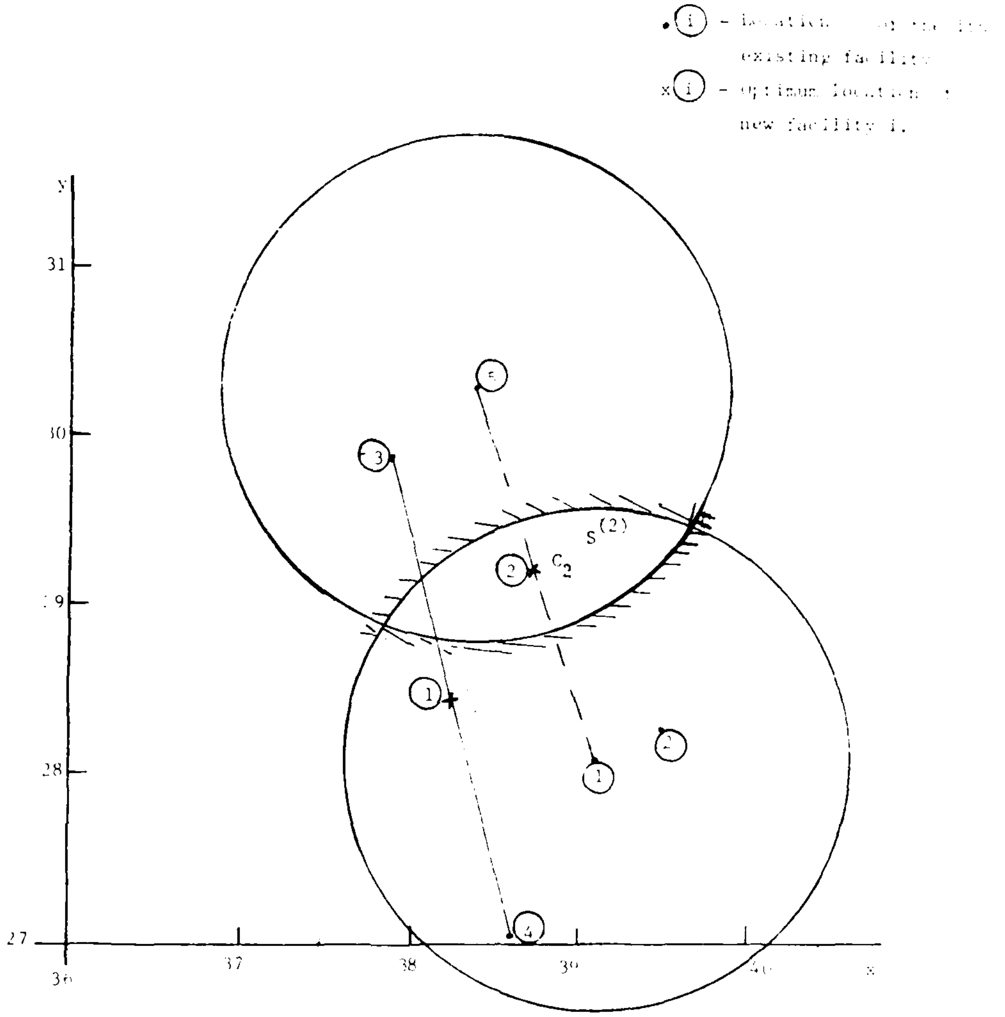


FIGURE 1. Illustration of solution set for example 1.

The Revised Algorithm

In summary the revised algorithm operates in two phases:

(i) Use Phase 1 (algorithm in Section 3) with $\epsilon = 10^{-4}$ to get to the neighborhood of the solution and to identify the functions that are inactive at the solution.

(ii) Continue the algorithm by using Phase 2 where the inactive functions are excluded from any further consideration.

EXAMPLE 2: In this case $n = 3$, $m = 5$.

i	1	2	3	4	5
a_i	0	2	5	7	8
b_i	0	8	4	6	2

$$W = \begin{bmatrix} 6 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 3 & 4 \\ 0 & 5 & 2 & 0 & 2 \end{bmatrix}$$

$$v_{12} = 0, \quad v_{13} = 2, \quad v_{23} = 1.$$

The results obtained by using Phase 1 of the algorithm are summarized below. It can be seen that only functions f_{32} and f_{35} define the minimax function at the solution point, and $\lambda_{ij} \to 0$, for all (i, j) except λ_{32} and λ_{35}. Also

$$\mu_{lk} \to 0, \ 1 \leqslant l < k \leqslant 3.$$

Results for Example 2 using Phase 1 of the Algorithm

	Number of Iterations					
	1	10	20	30	40	59
$F(\bar{x}, \bar{y})$	13.0004	12.1901	12.1319	12.1242	12.1225	12.1219
$F_l(\bar{x}, \bar{y})$	09.0075	11.7768	12.0475	12.0996	12.1143	12.1206

Values of (λ_{ij}), (μ_{lk}), (f_{ij}), (g_{lk}) and $(\bar{x}, \bar{y})$ after 59 iterations:

$$\Lambda = (\lambda_{ij}) = \begin{bmatrix} 0.0002 & 0. & 0. & 0. & 0. \\ 0. & 0. & 0. & 0. & 0. \\ 0. & 0.2855 & 0. & 0. & 0.7136 \end{bmatrix}, \ \mu_{12} = \mu_{13} = \mu_{23} = 0.$$

$$(f_{ij}(\bar{x}, \bar{y})) = \begin{bmatrix} 10.9488 & 6.5178 & 9.4821 & 0. & 0. \\ 0. & 0. & 2.5873 & 7.0682 & 07.0682 \\ 0. & 12.1216 & 5.2442 & 0. & 12.1219 \end{bmatrix}, \ g_{12} = 0, \ g_{13} = 10.9495, \ g_{23} = 4.6361$$

$$\bar{x}_1 = 0.92850, \ \bar{x}_2 = 7.57143, \ \bar{x}_3 = 3.71311$$

$$\bar{y}_1 = 1.57092, \ \bar{y}_2 = 3.71429, \ \bar{y}_3 = 6.28460.$$

Starting from the results obtained at the end of the 59 iterations of Phase 1 and using Phase 2, the algorithm reached the *exact* solution to the problem in *two* additional iterations. The final results obtained are summarized below. The method required 0.85 sec CPU time to reach the final results shown.

Final Results for Example 2: $F(x^*, y^*) = F_l(x^*, y^*) = 12.1218305$, $\lambda^*_{ij} = 0.$, except $\lambda^*_{32} = 0.28571$, $\lambda^*_{35} = 0.71429$, $\mu^*_{lk} = 0.$, $\forall (l, k)$.

$$(f_{ij}(x^*, y^*)) = \begin{bmatrix} 10.949 & 6.5177 & 9.4821 & 0. & 0. \\ 0. & 0. & 2.5873 & 7.06818 & 07.06818 \\ 0. & 12.1218 & 5.2450 & 0. & 12.1218 \end{bmatrix}$$

$$g_{12}(x^*, y^*) = 0, \quad g_{13}(x^*, y^*) = 10.953, \quad g_{23}(x^*, y^*) = 4.6357$$

$$\left.\begin{array}{ll} x_1^* = 0.92850 & y_1^* = 1.57091 \\ x_2^* = 7.57143 & y_2^* = 3.71429 \\ x_3^* = 3.71429 & y_3^* = 6.28571 \end{array}\right\} \text{Exact solution.}$$

Since f_{32} and f_{35} are the only functions defining the minimax solution and both of them depend only on (x_3, y_3) the values of (x_1^*, y_1^*) and (x_2^*, y_2^*) are not unique, but the value of (x_3^*, y_3^*) is unique. In fact (x_1^*, y_1^*) in any point in the set

$$S^{(1)} = \bigcap_{j=1}^{6} S_j^{(1)}$$

and (x_2^*, y_2^*) is any point in the set

$$S^{(2)} = \bigcap_{j=1}^{6} S_j^{(2)}$$

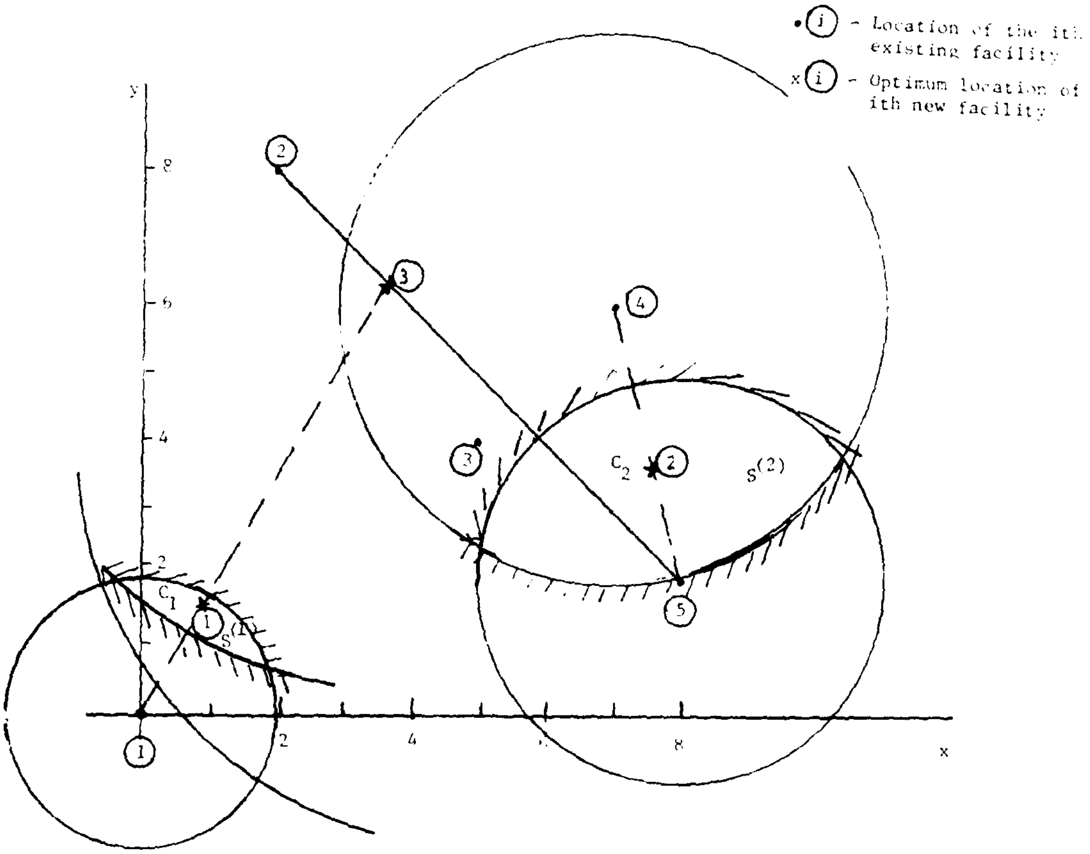
where

(15) $$S_j^{(i)} = \{(x, y) | w_{ij}(x - a_j)^2 + (y - b_j)^2]^{1/2} \leqslant F(x^*, y^*)\},$$

$$i = 1, 2, j = 1, 2, \ldots, 5$$

$$S_6^{(i)} = \{(x, y) | v_{13}(x - x_3^*)^2 + (y - y_3^*)^2]^{1/2} \leqslant F(x^*, y^*)\}, \; i = 1, 2.$$

The solution sets are illustrated in Figure 2.



•(i) - Location of the ith existing facility
x(i) - Optimum location of ith new facility
y
8
6
4
2
2
3
4
3
C2
2
S(2)
C1
1
S(1)
5
1
2
4
8
x

FIGURE 2. Illustration of solution sets for example 2.

As in Example 1, it would be appropriate to choose the position of the first and the second new facilities such that the function

$$F_{1,2}(x_1, y_1, x_2, y_2) = \max_{\substack{1 \leqslant i \leqslant 2 \\ 1 \leqslant j \leqslant 5 \\ 1 \leqslant l \leqslant 2}} \{f_{ij}(x_i, y_i), g_{l3}(x_l, y_l, x_3^*, y_3^*), g_{12}(x_1, y_1, x_2, y_2)\}$$

is minimized, subject to the conditions $(x_1, y_1) \in S^{(1)}$ and $(x_2, y_2) \in S^{(2)}$. Since $v_{12} = 0$ function g_{12} can be excluded in defining function $F_{1,2}$. The optimum solution to this problem is such that (x_1^*, y_1^*) is unique and is that obtained by using the present algorithm (Point C_1 in Figure 2), and (x_2^*, y_2^*) in any point in the set $S^{(2)}$.

Again it would be appropriate to choose the position of the second new facility such that the function

$$F_2(x_2, y_2) = \max_{1 \leqslant j \leqslant 5} \{f_{2j}(x_2, y_2), g_{21}(x_2, y_2, x_1^*, y_1^*), g_{23}(x_2, y_2, x_3^*, y_3^*)\}$$

is minimized. The optimum solution to this problem occurs at point C_2 in Figure 2 and is that obtained by the present algorithm.

5. CONCLUSIONS

An algorithm for the minimax facility location problem using Euclidean distances was proposed. Although no proof of convergence of the algorithm is available, for all examples considered, the algorithm converged to a minimax solution. Since there is no line search in the algorithm it follows that one iteration is the same as one function evaluation.

6. ACKNOWLEDGMENT

The author wishes to thank P. Lafoyiannis for programming the present algorithm, and Paul Calamai for the useful criticisms. Furthermore, the author wishes to thank the referee for his valuable comments. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Charalambous, C., "Acceleration of the Least *p*th Algorithm for Minimax Optimization with Engineering Applications," Mathematical Programming, *17*, 270-297 (1979).

[2] Charalambous, C. and A.R. Conn, "An Efficient Method to Solve the Minimax Problem Directly," SIAM Journal of Numerical Analysis, *15*, 162-187 (1978).

[3] Chatelon, J.A., D.W. Hearn and T.J. Lowe, "A Subgradient Algorithm for Certain Minimax and Minisum Problems," Mathematical Programming, *15*, 130-145 (1978).

[4] Dem'yanov V.F. and V.N. Malozemov, "Introduction to Minimax," (John Wiley & Sons, New York, N.Y. 1974).

[5] Drezner Z. and G.O. Wesolowsky, "A New Method for the Multifacility Minimax Location Problem," The Journal of the Operational Research Society, *29*, 1095-1101 (1978).

[6] Dutta S.R.K. and M. Vidyasagar, "New Algorithms for Constrained Minimax Optimization," Mathematical Programming, *13*, 140-155 (1977).

[7] Elzinga, J., D.W. Hearn and W.D. Randolph, "Minimax Multifacility Location with Euclidean Distances," Transportation Science, *10*, 321-336 (1976).

[8] Francis, R.L. and J.A. White, "Facility Layout and Location: An Analytic Approach," (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974).

[9] Love, R.F., "Locating Facilities in Three-Dimensional Space by Convex Programming," Naval Research Logistics Quarterly, *16*, 503-516 (1969).

[10] Love, R.F., G.O. Wesolowsky and S.A. Kraemer, "A Multifacility Minimax Location Method for Euclidean Distances," International Journal of Production Research, *11*, 37-45 (1973).

[11] Madsen, K., "An Algorithm for Minimax Solution of Overdetermined Systems of Non-Linear Equations," Journal of the Institute of Mathematics and its Applications, *16*, 321-328 (1975).

[12] White, J.A., "A Quadratic Facility Location Problem," AIIE Transactions, *3*, 156-157 (1971).

COUNTEREXAMPLES TO OPTIMAL PERMUTATION SCHEDULES FOR CERTAIN FLOW SHOP PROBLEMS

S. S. Panwalkar, M. L. Smith

Department of Industrial Engineering
Texas Tech University
Lubbock, Texas

C. R. Woollam

Department of Management
The University of Tennessee
Knoxville, Tennessee

ABSTRACT

It is well known that a minimal makespan permutation sequence exists for the $n \times 3$ flow shop problem and for the $n \times m$ flow shop problem with no in-process waiting when processing times for both types of problems are positive. It is shown in this paper that when the assumption of positive processing times is relaxed to include nonnegative processing times, optimality of permutation schedules cannot be guaranteed.

1. INTRODUCTION

Consider the n job-m machine flow shop sequencing problem in which processing times are nonnegative. In the following we will show that a permutation schedule may not be optimal for the classical flow shop problem involving three machines and for the $n \times m$ flow shop problem with the no in-process waiting constraint. We will use the 4 × 3 problem data shown in Table 1 and the nonpermutation schedule P defined in Table 2. Note that job 2 does not require processing on machine B.

TABLE 1 — *Processing Time Matrix*

Job	Machine		
	A	B	C
1	1	6	3
2	2	0	4
3	4	1	3
4	2	3	1

TABLE 2 — *Nonpermutation Schedule P*

Machine	Job Order
A	1,2,3,4
B	1,3,4
C	2,1,3,4

2. THREE MACHINE FLOW SHOP PROBLEM

In [8], Johnson proved the optimality of a permutation schedule for the $n \times 2$ problem under the assumption of positive processing times. He then extended the results to the $n \times 3$ problem and proved that an optimal permutation schedule exists. A number of researchers [3, 9, 1-p.9, 2-p.136, 4-p.84, 5-p.343, 6-p.201] have since relaxed the assumption of positive processing times to nonnegative ones. It is easy to verify that for the problem in Table 1, an optimal permutation schedule has a makespan of 16 units while the nonpermutation schedule P defined above has a makespan equal to 14 units.

3. FLOW SHOP PROBLEM WITH NO IN-PROCESS WAITING

We now consider the $n \times m$ flow shop sequencing problem with no in-process waiting allowed [10, 11]. In [11], Wismer considers nonnegative processing times. However, he allowed only permutation schedules. In [2], Baker recognized the fact that a nonpermutation schedule may be optimal when processing times are nonnegative. Gupta [7], on the other hand, has proved (Theorem 1) that even when the processing times are nonnegative only permutation schedules are feasible. The example in Table 1 is a counterexample to Gupta's theorem. For the no waiting problem, the best permutation sequence has a makespan of 17 as opposed to sequence P which has a makespan of 15. It may be noted that in both cases above, the minimum problem size needed to obtain a better nonpermutation schedule is 3×3.

REFERENCES

[1] Ashour, S., *Sequencing Theory* (Springer-Verlag, New York, N.Y., 1972).

[2] Baker, K.R., *Introduction to Sequencing and Scheduling* (John Wiley & Sons, Inc., New York, N.Y., 1974).

[3] Burns, F. and J. Rooker, "A Special Case of the 3 × n Flow-Shop Problem," Naval Research Logistics Quarterly, *22*, 811-817 (1975).

[4] Conway, R.W., W.L. Maxwell and L.W. Miller, *Theory of Scheduling* (Addison-Wesley Publishing Co., Reading, Mass. 1967).

[5] Eiselt, H.A. and H. Von Frajer, Editors, *Operations Research Handbook* (Walter de Gruyter and Co., Berlin, 1977).

[6] Fabricki, W.J., P.M. Ghare and P.E. Torgersen, *Industrial Operations Research* (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1972).

[7] Gupta, J.N.D., "Optimal Flowshop with No Intermediate Storage Space," Naval Research Logistics Quarterly, *23*, 235-243 (1976).

[8] Johnson, S.M., "Optimal Two- and Three-State Production Schedules with Setup Times Included," Naval Research Logistics Quarterly, *1*, 61-68 (1954).

[9] Lomnicki, Z.A., "A Branch-and-Bound Algorithm for the Exact Solution of the Three-Machine Scheduling Problem," Operational Research Quarterly, *16*, 89-100 (1965).

[10] Reddi, S.S. and C.V. Ramamoorthy, "On the Flow Shop Sequencing Problem with No Wait in Process," Operational Research Quarterly, *23*, 323-331 (1972).

[11] Wismer, D.A., "Solution of the Flowshop Scheduling Problem with No Intermediate Queues," Operations Research, *20*, 689-697 (1972).

A NOTE ON A MAXIMIN DISPOSAL POLICY UNDER NWUE PRICING*

Manish C. Bhattacharjee

Indian Institute of Management
Calcutta, India

ABSTRACT

For the classical disposal model for selling an asset with unknown price distribution which is NWUE (new worse than used in expectation) with a given finite mean price, this note derives a policy which is maximin. The gain in using the maximin policy relative to the option of selling right away is convex decreasing in the continuation cost to mean price ratio. The relevant results of Derman, Lieberman and Ross also follow as a consequence of our analysis. Our theorem provides a practical justification of their main result on the cutoff bid for the disposal model subject to NWUE pricing.

1. INTRODUCTION

Consider an indivisible asset for which offers come in sequentially, with a continuation cost $c > 0$ for each day the bid is not accepted. When the successive offers are independent identically distributed with a distribution F, this classic disposal model has been reconsidered by Derman, Lieberman and Ross [3] in an adaptive setting and when F is NWUE (new worse than used in expectation). While a complete solution is given in the adaptive case, their main result in the other case provides a lower bound on the optimal cutoff bid which, except for implying a corresponding lower bound on the optimal return (viz. Theorem 1 and Proposition 2 in [3]), is of limited practical value if F is NWUE but unknown.

The purpose of this note is to show that when the pricing is NWUE with a given mean price but is otherwise unknown, there is a maximin disposal policy determined by the lower bound for the cutoff bid given in [3]. As a by-product of our analysis, the Derman-Lieberman-Ross results on the cutoff bid also follow directly without invoking the ordering relationship among distributions defined through integrals of increasing convex functions as considered in [3].

2. MAXIMIN POLICY UNDER NWUE PRICING

Let $\bar{F} = 1 - F$. For the classic disposal model [2], [3], with F known, recall there is an optimal policy maximizing expected return—which accepts offer x if and only if $x \geqslant x_F$, and has return $(x_F + c)$, where the optimal cutoff bid x_F is given by

*This research was supported by the Center for Management Development Studies at the Indian Institute of Management, Calcutta, India under research project 441/CMDS-APRP-I

$$(2.1) \qquad x_F = \inf\left\{z : z \geqslant \left[\int_z^\infty y\, dF - c\right] / \bar{F}(z)\right\}$$

$$= \inf\{z : c \geqslant E_F(X - z)^+\},$$

$X \geqslant 0$ being distributed as F. Let $x_{\exp}$ denote the optimal cutoff bid for an exponential price distribution with the same mean as that of F, this distribution being henceforth abbreviated as 'exp'. Then

$$(2.2) \qquad x_{\exp} = -m \log (c/m), \quad \text{where } m = \int_0^\infty \bar{F}(y)\, dy > c.$$

Let

$$(2.3) \qquad L_F(x) = E \max(X, x - c) - x.$$

Note $L_F(c + x) = E(X - x)^+ - c$; thus (2.1), when F is continuous, implies $L_F(c + x_F) = 0$. Also, $L_F(x)$ decreases in x; this follows by noting $c + L_F(c + x) = \int_x^\infty \bar{F}(y)\, dy$.

Let π denote any policy (including randomized ones with past memory) and $R(\pi, F)$ its return. For any x, let $\pi(x)$ be the (stationary nonrandomized) policy which sells as soon as a bid of amount x or more is received. For any x such that $F(x) < 1$, the return $R(x,F)$ of the policy $\pi(x)$ is:

$$(2.4) \qquad R(x,F) = \sum_{n=1}^\infty \{E(X \mid X \geqslant x) - (n-1)c\}\, F^{n-1}(x)\, \bar{F}(x)$$

$$= E_F(X \mid X \geqslant x) - \frac{c\,F(x)}{1 - F(x)}$$

$$= x + [E_F(X - x)^+ - cF(x)]/\bar{F}(x)$$

$$= x + c + [L_F(c + x)/\bar{F}(x)].$$

Now suppose the pricing distribution F with mean $m < \infty$ has the NWUE property [1] defined by

$$\int_x^\infty \bar{F}(y)\, dy \geqslant m\, \bar{F}(x)$$

i.e., $\inf_{x \geqslant 0} E_F(X - x \mid X \geqslant x) = E_F X$. Then we have the following:

THEOREM: Suppose the price distribution F is NWUE and the continuation cost $c < m = E_F X < \infty$. If we only know the mean m (and not F), then the policy which sells as soon as the offered price is $x_{\exp}$ or more is maximin.

To prove the theorem, we will use the following generalization of a result (lemma 6.4, p. 112) in [1], a direct application of which yields Proposition 2 and Theorem 1 of [3].

LEMMA: If F is NWUE with mean $m < \infty$ and $\phi(y)$ is nondecreasing on $[0, \infty)$, then

$$\int_0^\infty \phi(y)\, \bar{F}(y)\, dy \geqslant \int_0^\infty \phi(y)\, e^{-y/m}\, dy.$$

If F is NBUE (new better than used in expectation), the inequality is reversed.

PROOF: Let Y be a random variable distributed as $TF(x) \overset{\text{def}}{=} m^{-1} \int_0^x \bar{F}(y)\, dy$ and let Z be exponential with mean m. Now F is NWUE implies the inequality $\int_x^\infty \bar{F}(y)\, dy \geqslant m e^{-x/m}$ (viz., [1], p. 187), i.e., Z is stochastically smaller than Y. Hence,

$$\int_0^\infty \phi(y) \bar{F}(y)\, dy = m \int_0^\infty \phi(y)\, TF(dy)$$

$$= m\, E\phi(Y) \geqslant m\, E\phi(Z) = \int_0^\infty \phi(y)\, e^{-y/m}\, dy.$$

The NBUE case $(E_F(X - x | X \geqslant x) \leqslant E_F X)$ follows by reversing all inequalities.

PROOF of Theorem: For any $x \geqslant 0$, choose ϕ, in the lemma, as the indicator of $[x, \infty)$ to conclude

(2.5) $$c + L_F(c + x) = \int_x^\infty \bar{F}(y)\, dy \geqslant \int_x^\infty e^{-y/m}\, dy = c + L_{\exp}(c + x),$$

when F is NWUE. Thus, $L_F(c + x) \geqslant L_{\exp}(c + x)$. This with (2.1) implies that $x_f \geqslant x_{\exp}$, as in Derman, Lieberman and Ross [3]. Hence, when F is NWUE, by (2.4) we have

(2.6) $$R(x_{\exp}, F) \geqslant c + x_{\exp},$$

since $L_F(c + x_{\exp}) \geqslant L_{\exp}(c + x_{\exp}) = 0$, where the inequality is due to the NWUE hypothesis and the last equality holds by continuity of the exponential distribution. Also, for any F,

(2.7) $$\sup_\pi R(\pi, F) = c + x_F = R(x_F, F),$$

since the policy $\pi(x_F)$ has the maximal return for a given price distribution F. Hence, using (2.5), (2.6) and (2.7), and $\inf_F$ denoting infimum over all NWUE distributions F with a given mean m, we have

$$\begin{aligned} c + x_{\exp} \leqslant \inf_F R(x_{\exp}, F) &\leqslant \sup_x \inf_F R(x, F) \\ &\leqslant \sup_\pi \inf_F R(\pi, F) \\ &\leqslant \inf_F \sup_\pi R(\pi, F) \\ &\leqslant \sup_\pi R(\pi, \exp) \\ &= R(x_{\exp}, \exp) = c + x_{\exp}. \end{aligned}$$

Thus, $R(x_{\exp}, \exp) = \sup_\pi \inf_F R(\pi, F)$ and the policy $\pi(x_{\exp})$ is maximin, i.e., it maximizes the reward from the worst possible NWUE law with given mean.

REMARKS:

1. Note, (2.5) together with (2.1) implies Proposition 2 of [3], by arguments paralleling those leading to (2.6). Likewise, the main result (Theorem 1) of [3] for NWUE pricing is contained in the proof of our Theorem.

2. The maximin policy behaves as if the price distribution, with known mean m, is exponential. Its relative gain compared to selling right away is

$$m^{-1} R(x_{\exp}, \exp) - 1 = -(1 - \alpha) - \log_e \alpha > 0,$$

where $\alpha = c/m$, the continuation cost to mean price ratio; $0 < \alpha < 1$. The relative gain increases as α decreases.

3. Suppose the price distribution F is arbitrary but strictly increasing and let $\xi_{\frac{1}{2}}$ be the median price. Then we will show:

$$x_f \geq (\xi_{\frac{1}{2}} - 2c)^{+}. \tag{2.8}$$

Take $c < \frac{1}{2}\xi_{\frac{1}{2}}$. If (2.8) does not hold, then $x_f < (\xi_{\frac{1}{2}} - 2c)$ and using (2.4) and (2.7), we have

$$c + x_f \geq R(2c + x_f, F) \geq 2c + x_f - c\{F(2c + x_f)/\bar{F}(2c + x_f)\} > c + x_f,$$

a contradiction. When the price distribution is NWUE, a bound stronger than (2.8) actually holds. To see this, note that if F is NWUE with mean m, then using (2.4) we get

$$\begin{aligned} R(x,F) &= x + E_f(X - x \mid X \geq x) - c\{F(x)/\bar{F}(x)\} \\ &\geq x + m - c\{F(x)/\bar{F}(x)\} \end{aligned} \tag{2.9}$$

for all x such that $F(x) < 1$. Accordingly,

$$c + x_f = R(x_f, F) \geq R(\xi_{\frac{1}{2}}, F) \geq \xi_{\frac{1}{2}} + m - c,$$

where the first inequality is due to (2.7) and the next one follows from (2.9). Hence, $x_f \geq m + \xi_{\frac{1}{2}} - 2c > \xi_{\frac{1}{2}} - 2c$ and (2.8) holds afortiori. Since x_f is nonnegative and $a > b$ implies $a^+ \geq b^+$, the resulting inequality $x_f \geq (m + \xi_{\frac{1}{2}} - 2c)^+$ is a sharpening of (2.8).

ACKNOWLEDGMENT

Thanks are due to the referee for helpful comments.

Note added in proof: Bengt Klefsjö, in a private communication, has recently pointed out to the author that our results (main theorem and remarks) remain valid for the broader class of HNWUE (harmonic new worse than used in expectation) price distributions. The classes HNWUE (HNBUE) which are less well known, were introduced by Rolski [5] and further studied by Klefsjö [4], are strictly bigger than NWUE (NBUE). A life distribution F with mean m is said to be HNWUE (HNBUE) if

$$\int_x^\infty \bar{F}(y)\,dy \underset{(\leq)}{\geq} m e^{-x/m}. \tag{2.10}$$

The reason for the name HNWUE (HNBUE) derives from the fact that (2.10) is equivalent [4] to

$$\left[\frac{1}{x}\int_0^x \{E_F(X - y \mid X \geq y)\}^{-1}\,dy\right]^{-1} \underset{(\leq)}{\geq} m = E_F X.$$

It can be easily seen that the Lemma remains true under HNWUE (HNBUE) hypothesis and hence our results carry over to HNWUE price distributions.

REFERENCES

[1] Barlow, R. and F. Proschan, *Statistical Theory of Reliability and Life Testing Probability Models* (Holt, Rinehart and Winston, New York, N.Y., 1975).

[2] Chow, Y.S. and H. Robbins, "A Martingale Systems Theorem and Applications," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 93-104 (1961).

[3] Derman, C., G.J. Lieberman and S.M. Ross, "Adaptive Disposal Models," Naval Research Logistics Quarterly, *26*, 33-40 (1979).

[4] Klefsjö, B., "Some Properties of the HNBUE and HNWUE classes of Life Distributions," *Statistical Research Report* #1980-8, University of Umeå, Sweden (1980).

[5] Rolski, T., "Mean Residual Life," *Proceedings of the 40th session*, Bulletin of the Int'l Stat. Inst., Voorburg, Netherlands, *4*, 266-270 (1975).

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

JUNE 1981
VOL. 28, NO. 2
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217

END